Poverty indicators at NUTS-2 level by Small Area Estimation

Eurostat Grant: Action plan for EU-SILC improvements

Methodological part

Version 26/01/2018

Table of Contents

List of figures

List of tables

1 Introduction

Statistics Belgium plans to use Small Area Estimation methods for the estimation of poverty indicators at NUTS-2 level starting from SILC 2018. The plan is to design a model based on an extensive use of administrative data, in order to deliver reliable and stable results at NUTS 2 level, as expected by Eurostat. A part of the grant, according to the methodology, is to study the administrative data, its relations with poverty indicators, and the possibility to use it in a suitable model at NUTS 2 level.

Small area estimation (SAE) is usually a useful theory to provide results for domains "for which direct estimates of adequate precision cannot be produced."[1] This theory involves using data from an entire sample for small domains (like a NUTS-2) estimation, according to auxiliary variables for each domain. The major interest of SAE is to have a larger sample size for each NUTS-2 and thus, to provide more stable and accurate results. As we will see in section 2, NUTS-2 in Belgium can be considered as small area, because of a small sample size.

A « small area » model requires auxiliary data. For a given NUTS-2 and a given year, we need:

- Sampling data from another NUTS-2
- Sampling data from previous year
- Population data from the NUTS-2

Here, we use these two types of data:

- Sampling data refer to current and previous years. These data are used to design the model
- Population data refer to the current year for each NUTS-2. These data are used to provided estimators for each NUTS-2

Our first attempt was to use the "synthetic estimator" introduced by Pascal Ardilly[2], for its simplicity and clarity. It is a calibration method of the entire database on NUTS-2 margins for one or many auxiliary variables (see previous report). We saw that this model provides stable results: value and rank of AROP rates by NUTS-2 seem to be, thanks to this method, stable over time. Despite these encouraging results, synthetic estimators are "too" stable, and cannot reflect real variations at NUTS-2 level. And, because of the regional concordance needed to our result, we obtain some inconsistent results between direct and synthetic estimation. So synthetic estimator was discarded and we investigate now for more accurate models.

In this report, we present our final work about using Small Area Estimation (SAE) methods, using the theory of mixed model (see section 3). We describe a methodological way to validate the relevance of SAE methods in our case. We give some evidences that SAE methods will upgrade the precision and the stability of poverty indicators at NUTS-2 level. Furthermore, we apply a non-linear mixed model to our data to provide poverty indicators (AROP, LWI, SMD, AROPE) for SILC 2011 to SILC 2016.

[1] Rao, J. N. (2003). Small area estimation. John Wiley & Sons, Inc.

[2] Ardilly, Pascal (2013). "Regional estimates of poverty indicators based on a calibration technique", Eurostat Statistical working paper.

The first section presents our work on testing a Small Area Estimation model. We explain how to test the accuracy of SAE methods, and we apply these tests to our explanatory variables. The second section details the theory of Mix Model and the establishment of an EBLUP indicator. We explain how to adapt this theory into SAS and with national institute of statistics constraints the like regional concordance of the results. Section 3 presents our final result for SILC 2016, with an explanation of the trend from 2011 to 2016. We show that EBLUP estimation is more stable than direct estimation, and that the difference between direct and EBLUP estimation can be explained by the feature of the sample in some little areas. Finally, the last section explains our future actions to publish in the future the official poverty indicators at NUTS-2 level.

2 Justification of the small area estimation

Small area estimation (SAE) is usually a useful theory to provide results for domains "for which direct estimates of adequate precision cannot be produced."[3] This theory involves using data from an entire sample for small domains (like a NUTS-2) estimation, according to auxiliary variables for each domain. The major interest of SAE is to have a larger sample size for each NUTS-2 and thus, to provide more stable and accurate results.

In our case, SAE is accurate to provide poverty indicator at NUTS-2 in Belgium. Indeed, the SILC survey was planned to provide results at national level, and some NUTS-2 sampling sample are very small (for example, around 300 in the province of Luxembourg). As result, the AROP poverty indicator can have some high variations year per year (figure 1 next page). These results are unrealistic and cannot be published. We need to stabilize them.

Because of the low sampling size, the sample can have some different characteristics from the entire population. We have access to administrative data for the population, so we can check the difference between the sample and the population. We highlight 5 explanatory variables[4] provided by administrative data:

- Fiscal Income: we collect the fiscal income (N-2) for all households in Belgium. For each individual, we affect the fiscal income of his household
- Non-work income : we collect all the allocations (N-2) perceived by a household for a weak workforce (pension, unemployment, sickness)
- Age
- Gender
- Household size : we use four categories : 1,2,3 or 4 and more members in the household

[3] Rao, J. N. (2003). Small area estimation. John Wiley & Sons, Inc.

[4] Some other variables need to be analysed. For example, the owner/renter status could help to explain SMD or AROP, and we will have this status by administrative data for the entire population in the future.

Figure 1 : AROP indicator at NUTS-2 level for SILC 2011 to 2016

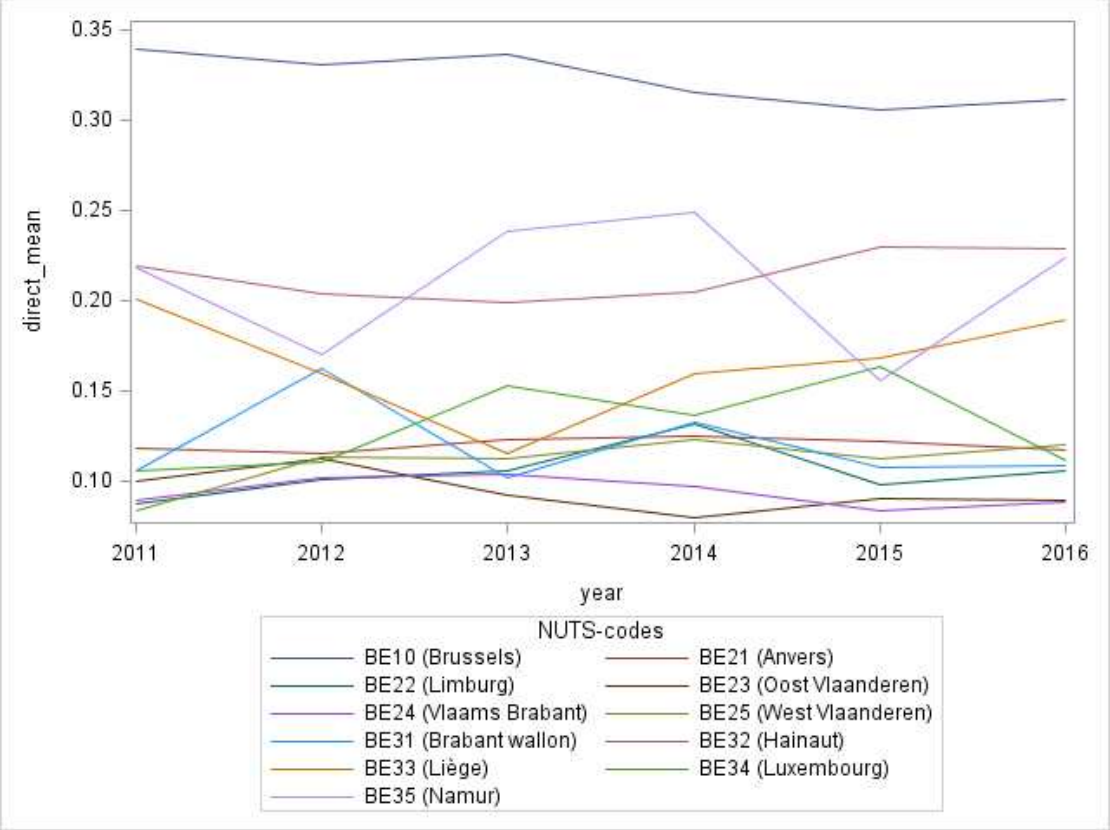


Table 1 (next page) shows the differences between the sample and the population, for Belgium and for each NUTS-2. There is no difference at national level, but some high differences at NUTS-2 level for the two fiscal variables. In the table, we highlight the variable when the ratio Sample/Population is higher than 5%.

In order to use these auxiliary variables, we need to prove that a model of our poverty indicators is relevant (otherwise, the estimator would be biased). We test four successive hypotheses for each poverty indicators:

1. The interest variable is correlated with the NUTS-2 variables : if not, the NUTS-2 analysis has no sense and providing results at national level is more relevant ;
2. The auxiliary variables are correlated with the interest variable : for each auxiliary variable and each interest variable, we have to check if the link is strong to justify the model ;
3. According to the auxiliary variables, the interest variable is not correlated with the NUTS-2 variables anymore ;
4. The link between the interest variable and the auxiliary variables is not altered by the NUTS-2 variables.

Table 1 : Mean of administrative data, for the sample and the population (Belgium and NUTS-2)

		Fiscal income			Non-work income		
NUTS	N Sample	Sample	Population	Ratio	Sample	Population	Ratio
Belgium	13681	22200	21800	1,02	11100	10700	1,04
BE10 (Brussels)	2365	17800	17200	1,03	9200	8400	1,10
BE21 (Antwerpen)	1735	23800	22600	1,05	12300	10400	1,18
BE22 (Limburg)	893	23400	22400	1,04	11400	11900	0,96
BE23 (Oost Vlaanderen)	1734	24500	23300	1,05	9500	10500	0,90
BE24 (Vlaams Brabant)	1226	24800	24600	1,01	10900	10900	1,00
BE25 (West Vlaanderen)	1582	21900	22700	0,96	10200	10900	0,94
BE31 (Brabant Wallon)	427	22900	23800	0,96	13100	11500	1,14
BE32 (Hainaut)	1754	20000	19600	1,02	11200	11800	0,95
BE33 (Liège)	1250	21400	20400	1,05	12600	11100	1,14
BE34 (Luxembourg)	311	22000	23600	0,93	13000	10100	1,29
BE35 (Namur)	404	20200	21200	0,95	11800	11100	1,06

Table 1 continued

	Age			Sex			Household size		
NUTS	Sample	Population	Ratio	Sample	Population	Ratio	Sample	Population	Ratio
Belgium	40,9	40,6	1,01	1,51	1,51	1,00	2,09	2,09	1,00
BE10	36,9	36,7	1,00	1,51	1,51	1,00	2,07	2,06	1,00
BE21	42,5	40,9	1,04	1,5	1,5	1,00	2,03	2,11	0,96
BE22	42,1	41,8	1,01	1,52	1,5	1,01	2,1	2,13	0,99
BE23	40,8	41,2	0,99	1,51	1,5	1,01	2,07	2,08	1,00
BE24	41,4	41,1	1,01	1,5	1,51	0,99	2,1	2,12	0,99
BE25	41,7	43,1	0,97	1,51	1,5	1,01	2,22	2,05	1,08
BE31	42,0	40,4	1,04	1,51	1,51	1,00	2,12	2,15	0,99
BE32	40,3	40,2	1,00	1,51	1,51	1,00	2,1	2,07	1,01
BE33	40,6	40,3	1,01	1,52	1,51	1,01	2,05	2,05	1,00
BE34	41,5	39,1	1,06	1,5	1,5	1,00	2,07	2,11	0,98
BE35	39,7	40,1	0,99	1,51	1,51	1,00	2,06	2,09	0,99

These four hypotheses are tested with logistic regressions, using three different models to explain the same variable (one of the poverty indicators). The three models differ according to their explanatory variables:

- Model A: only the NUTS-2 variables
- Model B: only the auxiliary variables
- Model C: the NUTS-2 variables and the auxiliary variables

We test hypothesis 1 through the A model. We test hypothesis 2 through the B model. We test hypothesis 3 by comparing model A to model C. Finally, we test hypothesis 4 by comparing model B to model C.

We test these hypotheses for 4 poverty indicators: AROP (at risk of poverty), LWI (low workforce intensity), SMD (severe material deprivation) and AROPE (at risk of poverty or social exclusion). In this report, we focus on AROP, because it is the main poverty indicator for SILC. Most of the conclusions remain identical for the 3 others poverty indicators and all the results are available in annex. If there is some difference, we discuss about them in the report. The following sub-sections test each hypothesis.

2.1 Link between the interest variable and the NUTS-2 variables

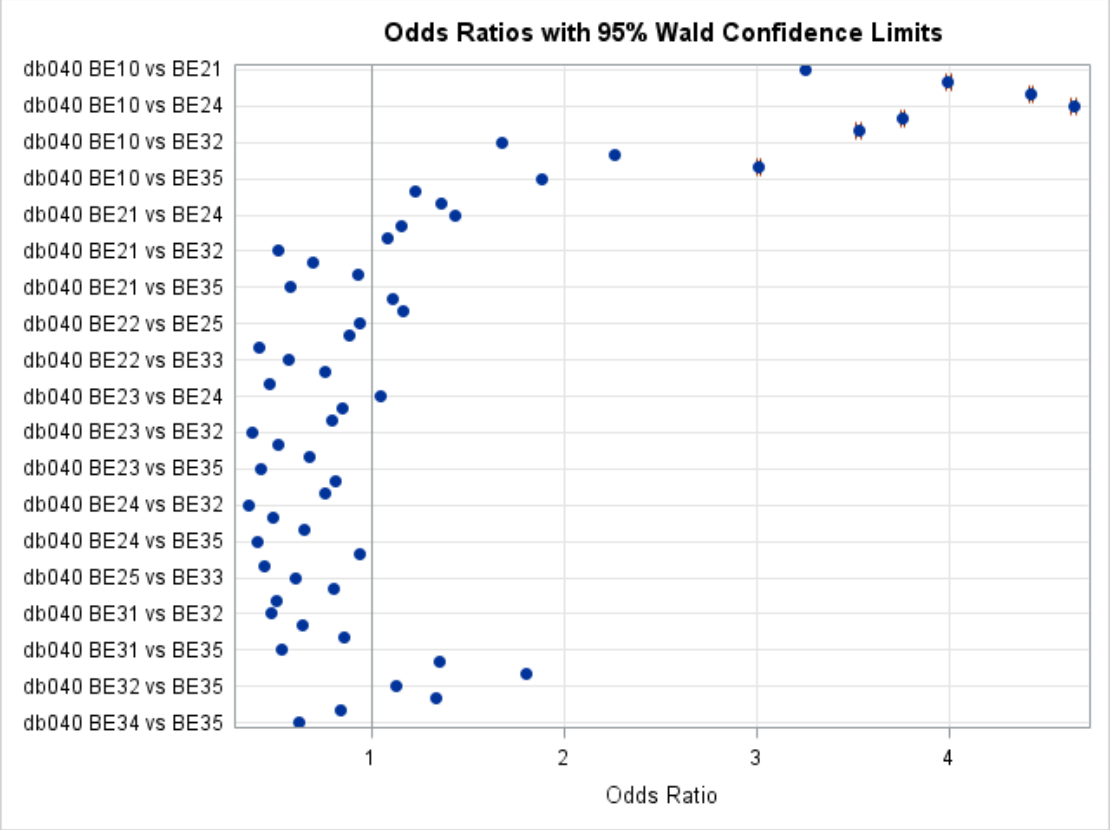
In Belgium, the several NUTS-2 have specific profile according to AROP (see table 2). A household who comes from one NUTS-2 has not the same "risk of AROP" than one coming from another NUTS-2, and we can quantify these risks with a logistic regression (model A). The same trend can be show for LWI, SMD and AROPE.

Table 2: AROP rate by NUTS-2 (individual level)

	AROP rate
BE10 (Brussels)	33 %
BE21 (Antwerpen)	12 %
BE22 (Limburg)	10 %
BE23 (Oost Vlaanderen)	11 %
BE24 (Vlaams Brabant)	10 %
BE25 (West Vlaanderen)	11 %
BE31 (Brabant Wallon)	16 %
BE32 (Hainaut)	20 %
BE33 (Liège)	16 %
BE34 (Luxembourg)	11 %
BE35 (Namur)	17 %

By studying odds ratios of this logistic regression, we emphasize initial differences between NUTS-2. Odds ratios are always calculated between two NUTS-2 and can be explained as follows: if an odds ratio is close to 1, then the two corresponding NUTS-2 have the same AROP profile ; if an odds ratio is far from 1, then the two corresponding NUTS-2 have different AROP profiles. The figure 2 (next page) shows the odds ratios by NUTS-2.

Figure 2 : Odds ratios by NUTS-2 for AROP rate (model A)



The table 3 summarizes the odds ratios by NUTS-2 (see also figure in annex for all the combinations). We notice that odds ratios are very high concerning Brussels and Namur. We will pay attention for these two NUTS-2 during the next steps.

Table 3: Summary of Odds ratios by NUTS-2 (AROP, model B)

NUTS-2	Mean	Standard deviation	Min	Max
BE10 (Brussels)	1,2235	0,780	0,515	3,251
BE21 (Antwerpen)	1,1533	1,040	0,418	4
BE22 (Limburg)	1,1592	1,192	0,378	4,431
BE23 (Oost Vlaanderen)	1,1763	1,268	0,36	4,651
BE24 (Vlaams Brabant)	1,0806	0,967	0,445	3,764
BE25 (West Vlaanderen)	1,0499	0,892	0,474	3,533
BE31 (Brabant Wallon)	0,8541	0,575	0,36	1,8
BE32 (Hainaut)	0,9273	0,565	0,486	2,259
BE33 (Liège)	1,1433	0,752	0,626	3,013
BE34 (Luxembourg)	0,7394	0,459	0,406	1,887
BE35 (Namur)	3,2463	1,035	1,674	4,651
Total	**1,2503**	**1,0779**	**0,36**	**4,651**

For the 3 other poverty indicators, we obtain the same results. Odds ratios are mostly different to 1 for all NUTS-2 (figure 16, 18, and 20 in annex).

We validate hypothesis 1: The interest variable is correlated with the NUTS-2 variables. Based on this result, we will check if adding auxiliary variables could bring odds ratio to 1.

2.2 Link between the auxiliary variables and the interest variable

Based on the fiscal income, the income quintiles are highly correlated to AROP, as reflected in table 4. The majority of households who are AROP belong to the first five income quintiles (IQ). This finding is consistent when we checked by NUTS-2 (see table 15 in annex). We also notice that some of the households in the 10 last IQ are AROP according to our survey. Further investigations are being carried, in order to understand this surprising result.

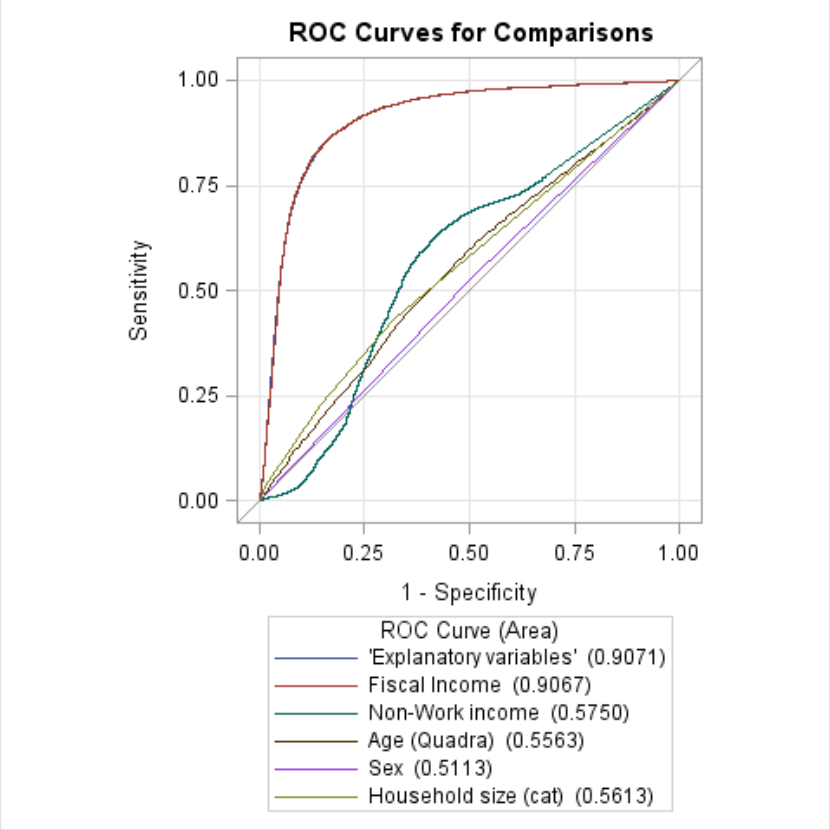
Table 4: AROP rate by income quintiles (individual level)

Income quintiles	AROP rate	Income quintiles	AROP rate
1	71 %	12	3 %
2	75 %	13	2 %
3	66 %	14	1 %
4	39 %	15	1 %
5	26 %	16	2 %
6	15 %	17	1 %
7	16 %	18	1 %
8	8 %	19	2 %
9	3 %	20	1 %
10	4 %	Not available	18 %
11	3 %		

When we look about all the auxiliary variables (not only the fiscal income), the logistic regression with AROP (model B), confirms the link. The concordance rate is 90%[5] with a ROC curves illustrated by the figure 3 (next page).

5 "A pair of observations with different observed responses is said to be *concordant* if the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value. If the observation with the lower ordered response value has a higher predicted mean score than the observation with the higher ordered response value, then the pair is *discordant*. If the pair is neither concordant nor discordant, it is a *tie*." (SAS, user guide, second edition)

Figure 3: Explanatory power of variables for AROP rate (model B)

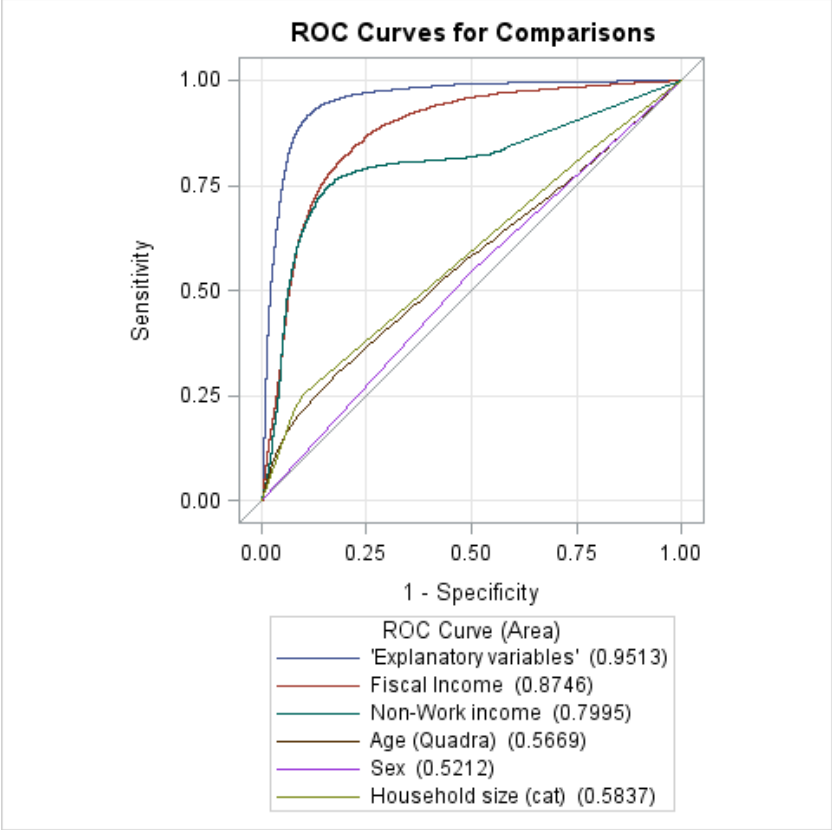


This ROC curve can be analyzed as follows: it is possible to obtain a small false negative rate (incorrectly conclude that a household is AROP) while obtaining a high true positive rate (correctly conclude that a household is AROP). With this model, it is possible to predict if a household is (or not) AROP according to his administrative profile with a high rate of success.

With this figure, we can notice that the fiscal income have the same explanatory power than all the auxiliary variables. We keep the other variables because they can improve the model quality for the other poverty indicators. For example, for LWI, fiscal income performs less and Non-Work Income has a better explanatory power (figure 4 next page).

We validate hypothesis 2: The auxiliary variable is correlated with the interest variable.

Figure 4: Explanatory power of variables for LWI rate



2.3 Link between interest and NUTS-2 variables, with regards to the auxiliary variables

We focus here on the difference between model A (only NUTS-2 variables) and model C (auxiliary variables and NUTS-2 variables). Between these two logistics models, the percentage of concordant predictions grows from 60% to 90%, so adding the auxiliary variables enhance the quality of the model.

Previously, we saw that odds ratios of the model A were highly different from 1. Now, with the model C, odds ratios of the NUTS-2 are closer to 1 (table 5 next page and figure 15 in annex).

Table 5: Part of Odds ratios significantly different to 1 (AROP, model C)

NUTS-2	Mean	Standard deviation	Min	Max
BE10 (Brussels)	1,6302	0,398	1,083	2,173
BE21 (Antwerpen)	1,0918	0,304	0,687	1,576
BE22 (Limburg)	1,0552	0,322	0,649	1,669
BE23 (Oost Vlaanderen)	1,0191	0,378	0,585	1,849
BE24 (Vlaams Brabant)	1,0187	0,518	0,498	2,173
BE25 (West Vlaanderen)	1,0006	0,448	0,536	2,02
BE31 (Brabant Wallon)	0,9312	0,289	0,696	1,556
BE32 (Hainaut)	0,8776	0,425	0,498	1,898
BE33 (Liège)	0,898	0,360	0,548	1,726
BE34 (Luxembourg)	1,3157	0,461	0,551	2,054
BE35 (Namur)	0,7495	0,218	0,521	1,132
Total	**1,0534**	**0,4314**	**0,498**	**2,173**

However, all correlations are not corrected. First, odds ratios are still different to 1, and odds ratios for Brussels remains very high. That is not a problem for Brussels, because Brussels is at the same time a NUTS-1 and a NUTS-2, so we will not use SAE for this province. For the other NUTS-2, there are some residual trends between AROP and the NUTS-2 and we still have to correct them.

We can see the same effect for LWI and AROPE (figure 17 and 21 in annex). The model reduces the correlation between the interest variable and the NUTS-2 variables, except for Brussels. For SMD, odds ratios remain very different from 1 (figure 19 in annex). The model performs less for this variable, and we have to study for more suitable variables.

We mostly validate hypothesis 3: According to the auxiliary variable, the interest variable is no more correlated with the NUTS-2 variables.

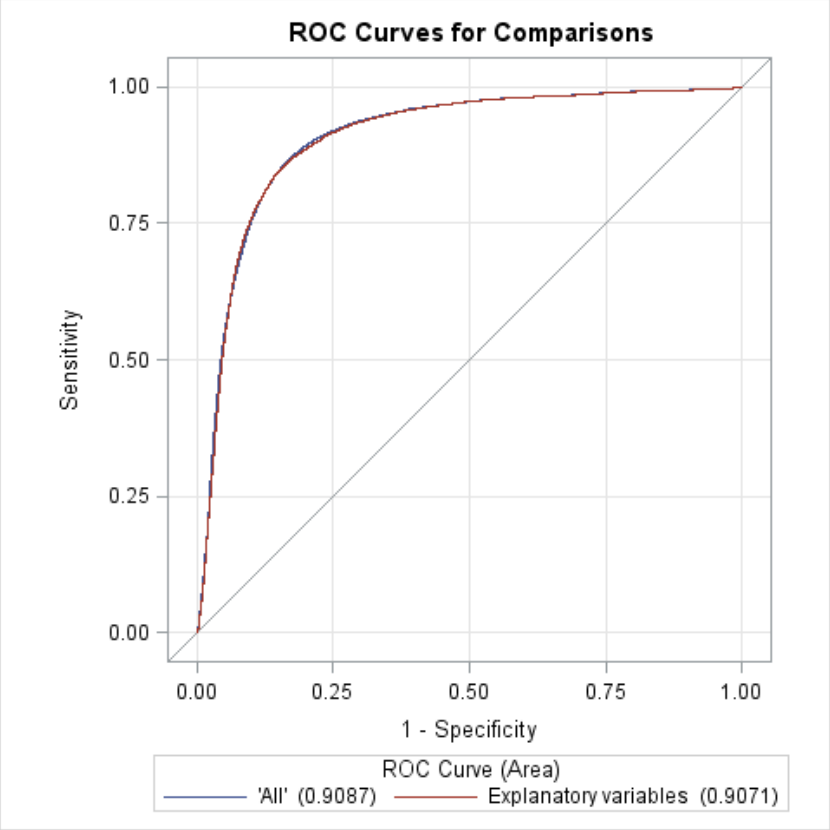
2.4 Link between interest and auxiliary variables, with regards to the NUTS-2 variables

When NUTS-2 variables are added to the model B, there is no sensible change of the results. About AROP, parameter estimates are almost identical, with or without inclusion of NUTS-2 variables (see table 6 next page). Despite this, in the model C, NUTS-2 parameters remains all significant (so some effects are only caused by the situation of the NUTS-2). Finally, when we compare ROC curves from model B and model C (figure 4 next page), we notice that including NUTS-2 variables changes nothing to the quality of the model.

Table 6 : Comparison between parameters of the models B and C (AROP, SILC 2016)

Parameter	Model B	Model C
Intercept	2,478	2,340
Fiscal income	-0,00024	-0,00024
Non-work income	1,15E-03	-6,22E-08
Age	-0,015	-0,0154
Age (square)	0,000153	0,000179
Sex	-0,0178	-0,0174
Household size (1)	-0,00044	-0,0399
Household size (2)	-0,1092	-0,1104
Household size (3)	0,051	0,0695

Figure 5 : ROC Curves for comparisons (AROP, model B versus model C)



This result is identical for LWI, SMD and AROPE (see figure 29, 30 and 31 in annex). For these 3 poverty indicators, NUTS-2 variables do not affect the model with our auxiliary variables.

We validate hypothesis 4: The link between the interest variable and the auxiliary variable is not altered by the NUTS-2 variables.

With these 4 hypotheses, we validate that a link exists between NUTS-2 and the poverty indicators, and that this link can be altered by our auxiliary variables. So, with a model of the poverty indicators, we will obtain estimators of poverty at NUTS-2 level without bias. The theory of this model is the purpose of the next section.

3 Mixed model theory and EBLUP estimator

To provide SAE estimators at NUTS-2 level, we need to have a model to explain a poverty indicator by some variables, available for the sample and for the entire population. After the justification of the model (section II), we present here the theoretical formalization of a mixed model with non-linear predictor.

First, we give a presentation of the mixed model. We restrain the presentation to linear predictor, to keep it simple. Then we explain the transition to the EBLUP estimator, from linear to non-linear predictor.

3.1 Mixed model theory

3.1.1 Presentation of data and notations

Let a sampling s drowned randomly into a population (universe) N divided into J domain. We assume that, for a year t, we have data for the T previous years. We have for each individual:

- A explained variable: $y_{i,j,t}$, available for the sampling ($i \in s$)
- Some explanatory variables: $x_{i,j,t}$
- A domain indicator: j
- A year indicator: t

We have also a population with the mean of each explanatory variable $\overline{X}_{j,t}$ by domain and by year.

3.1.2 Principles

A mixed model is called « mixed » because two types of effects are present in the model: some fixed effects, and some random effects. Fixed effects deal with control variables, here the auxiliary variables.

Random effects are not controlled by a variable, but these effects are predictable, for example here the domains and the years.

In the first instance, we only used sampling data. Here is the mixed model generic formula:

$$y_{i,j,t} = X_{i,j,t}\beta + \mu_j + \delta_t + \varepsilon_{i,j,t}$$

Vector β is the fixed effects for each explanatory variable. Vector μ is the random effect induced by each domain. In a same way, vector δ is the random effect induced by each year. Then $\varepsilon_{i,j,t}$ is the model residual.

We assume a Gaussian distribution for the random effects:

$$E(\varepsilon_{i,j,t}) = E(\mu_j) = E(\delta_t) = 0$$

$$V(\varepsilon_{i,j,t}) = \sigma_\varepsilon^2$$

$$V(\mu_j) = \sigma_\mu^2$$

$$V(\delta_t) = \sigma_\delta^2$$

Then we solve this program according to β, μ_j and δ_t by maximum likelihood estimation (in SAS, with PROC MIXED):

$$\min_{\beta,\mu_j,\delta_t} V(\hat{y}_{i,j,t} - y_{i,j,t})$$

$$s.t.\, E(\hat{y}_{i,j,t} - y_{i,j,t}) = 0$$

By applying the model, we can now provide estimation of the explained variable for an individual, according to his profile, for a given domain and a given year. To provide the domain estimator, we assume that N_j is big enough (it is always the case in our surveys) and we apply the model with the average of the explanatory variables:

$$\bar{Y}_{j,t} = \bar{X}_{j,t}\beta + \mu_j + \delta_t$$

3.2 Transition to BLUP estimator

In this part, to simplify, we use only one random effect (the domain effect). So, the random effect linked to the year is deleted. The explanation remains similar.

3.2.1 Composite estimator principle

Designed like this, the estimator $\bar{Y}_{j,t}$ takes into account all the data from the population, for a given domain. However, if the variance of the province random effect is strong, replacing the domain estimator by a national estimator is not relevant. In this case, it is more relevant to proceed to a "mix" [6]

[6] The word "mix" here is not related to the theory of the mixed model.

between data from the sampling (unbiased, because linked to the domain) and from the population (biased). It is almost the same case with the synthetic estimator from Ardilly (see previous interim report).

The direct estimator is assumed to be unbiased, because it depends only on the sampling data. On the other hand, it has a strong variance (it is the reason of using SAE). On the contrary, the synthetic estimator (based on a model at national level) is biased, but has a weaker variance. A mix between these two estimators, called composite estimator, offsets the advantages and the disadvantages of the two estimators. It involves doing a weighted average of them, according to the variance of the direct one. Stronger is the variance of the direct estimator, more we will use the synthetic estimator. On the contrary, weaker is the variance of the direct estimator (in case of a bigger sample size for a domain for example), more we will use the direct estimator.

BLUP estimator has exactly the same line of reasoning, except that the two estimators come from the mixed model. It consists on a share between applying the model with the sampling data and applying the model with the population data. If the variance of the random effect is strong (with respect to the residual variance), we will use mostly the data come from the sample. We are here talking about a regression estimator (that one based of sample data) and a synthetic estimator (that one based on population data).

3.2.2 Calculation of BLUP estimator

We come back with the equation of the mixed model estimator:

$$\bar{Y}_j = \bar{X}_j\beta + \mu_j$$

If the variance of the random effect linked to the domain $\left(\sigma_\mu^2\right)$ is weak, that means that μ_j are mostly near 0. So, for a given j, $\bar{Y}_j$ is near $\bar{X}_j\beta$ and applying the model to the population data is relevant. On the contrary, if the variance σ_μ^2 is strong, μ_j are mostly different than 0. So $\bar{Y}_j$ is different than $\bar{X}_j\beta$ and applying the model to the domain is more relevant.

We set the optimal proportion of the two estimators (regression and synthetics) by a ratio of the variances:

$$\gamma_j = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \frac{\sigma_\varepsilon^2}{n_j}}$$

This ratio γ_j is between 0 and 1, positively correlated with the random effects of the domains, positively correlated with the sample size of the domain and negatively correlated with the residual variance. According to this ratio, the BLUP is calculated by this formula:

$$\bar{Y}_j^E = \bar{X}_j\tilde{\beta} + \tilde{\mu}_j$$

With $\tilde{\mu}_j = \gamma_j \left(\bar{y}_j - \bar{x}_j \tilde{\beta}\right)$ and

$$\tilde{\beta}_j = \left(\sum_{i \in s_j} X_{i,j} X_{i,j}^T - \gamma_j \bar{x}_j \bar{x}_j^T \right)^{-1} \left(\sum_{i \in s_j} X_{i,j} Y_{i,j} - \gamma_j \bar{x}_j \bar{y}_j \right)$$

To compare, this is the "usual" formula for the vector of parameters β_j for a domain:

$$\beta_j = \left(\sum_{i=1}^{n_j} \frac{X_{i,j} \cdot X_{i,j}^T}{\sigma_\varepsilon^2} \right)^{-1} \left(\sum_{i=1}^{n_j} \frac{X_{i,j} \cdot Y_{i,j}}{\sigma_\varepsilon^2} \right)$$

Bigger is γ_j, more we will use data from the sample instead of those from the population. Indeed, in the formula of $\widetilde{\beta_j}$, data from the sample « cancel » those from the population. In the same time, parameter $\tilde{\mu}_j$ provides the estimation for the sample.

3.2.3 Transition to non-linear EBLUP estimator

Up to now, we calculated the BLUP estimator by assuming known variances. Solving this problem can be done by some optimization methods. But, frequently, we used the EBLUP estimator (E for Empirical) instead of BLUP estimator, for which unknown variances are replaced with their empirical estimation.

σ_μ^2 and σ_ε^2 are replaced by $\hat{\sigma}_\mu^2$ and $\hat{\sigma}_\varepsilon^2$. Everything else remains similar.

We know that there is a theoretical simplification in the current methodology presented above. Indeed, using a mix model as we did require a quantitative explained variable. Here, our explained variables (AROP, LWI, SMD, and AROPE) are qualitative. This preliminary work allows us to clarify the methods, formula and assumptions we can now use to provide a relevant model. To provide valid predictor, we use a logistic mix model with EBLUP estimators. The assumptions and the formalization are similar to those in this section.

Based on this theory, we can provide SAE results for SILC for each province in Belgium.

4 Poverty indicators at NUTS-2 level for SILC

In this last section, we present our application of SAE model and the first results for SILC at NUTS-2 level. All our methods are based on the mix model theory and are support by the stability of our results. We provide some explanation of our results, according to the difference at NUTS-2 level between the sample and the entire population means.

Here we present our specific methodology, with our technical and software constraints. Then we present and detail our results for SILC 2016 with a backward analysis from 2011 to 2015.

4.1 Software and macros

In Statistics Belgium, we use the SAS software to provide our statistics. We investigate tree main procedures to implement EBLUP estimator: PROC MIXED, PROC GLIMMIX and PROC NLMIXED.

PROC MIXED allows to implement easily EBLUP estimator. Indeed, these estimators are natively integrated into the procedure. Despite that, PROC MIXED only allows analysis on linear variable, so it is not suitable to our poverty indicators. We have to choose a non-linear procedure.

PROC GLIMMIX allows to implement non-linear estimator, with fixed and random effects. So, this procedure is suitable to our case. Despite that, EBLUP estimators are not directly calculated, so we have to regress global and individual model and after that to aggregate results our self to provide EBLUP estimators. Moreover, in our case, the convergence is rarely obtained, especially with random effects.

PROC NLMIXED is the freer procedure. Models, as random effects, are manually designed with all the parameters. Despite that, EBLUP estimators are not directly calculated neither the model residuals. Also, in our case, convergence is very difficult to obtain in a reasonable time.

In a simplified way, we will use PROC GLIMMIX and create our procedure to obtain EBLUP estimator by aggregation of global and individual. As we explain in the next section, we have to use PROC MIXED at the beginning of our procedure to skirt the lack of convergence.

4.2 Step by step procedure

In this sub-section, we detailed all the procedure we made to obtain EBLUP estimator. In each step, we illustrate and support our choices by some examples taken from SILC data.

4.2.1 Optimal proportion of the estimators

We first need to calculate the optimal proportion of the EBLUP estimators (based on the regression and synthetics models, see section III):

$$\gamma_j = \frac{\sigma_\mu^2}{\sigma_\mu^2 + \frac{\sigma_\varepsilon^2}{n_j}}$$

Because of convergence problems with the PROC GLIMMIX, we decide to estimate this proportion by a linear model. We assume that non-linear mix-model theory could be generalized, at this step, to a linear analysis when the sample size is enough.

For SILC 2016, the optimal proportion is between 0.5% et 5%, which means that the global model (all NUTS-2, all years) will be highly present into the EBLUP estimators (see table 7 next page). For years 2011 to 2015, these optimal proportions of estimator are quite the same.

Table 7 : Gamma by poverty indicator and NUTS-2, SILC 2016

NUTS-2	n (sample)	AROP	LWI	SMD	AROPE
BE10 (Brussels)	2365	4,09%	3,71%	5,41%	4,59%
BE21 (Antwerpen)	1735	3,03%	2,75%	4,02%	3,41%
BE22 (Limburg)	893	1,59%	1,44%	2,11%	1,78%
BE23 (Oost Vlaanderen)	1734	3,03%	2,75%	4,02%	3,41%
BE24 (Vlaams Brabant)	1226	2,16%	1,96%	2,88%	2,43%
BE25 (West Vlaanderen)	1582	2,77%	2,51%	3,68%	3,12%
BE31 (Brabant Wallon)	427	0,76%	0,69%	1,02%	0,86%
BE32 (Hainaut)	1754	3,07%	2,78%	4,07%	3,45%
BE33 (Liège)	1250	2,21%	2,00%	2,93%	2,48%
BE34 (Luxembourg)	311	0,56%	0,50%	0,75%	0,63%
BE35 (Namur)	404	0,72%	0,65%	0,97%	0,82%

4.2.2 Non-linear models

Next, we have to regress 12 logistics models.

First, we regress one logistic model for all the NUTS-2 and all the years (from N-2 to N), with NUTS-2 and year variables as random effects[7]. This model will provide synthetics estimation of parameters, including all the available data. This model is more stable because of the size of the sample, but can be biased because it aggregates several provinces and several years.

Then, we regress one logistic model by NUTS-2 for the year N. We call these the "regression" model, because they focus on one particular province, but they still are a model of the poverty indicator. Parameters from these models will be more instable, especially in a little province.

For each parameter, we apply the optimal proportion of the estimator between the synthetic and the regression estimation to obtain the EBLUP estimator. As we explained in the last sub-sub section, EBLUP estimators are mostly based on the synthetic model.

The following two pages show the list of parameters for AROP for SILC 2016. In the column Synthetic model, the parameters are always identical, except for the NUTS-2 parameter, because there is only one model. In the column Regression model, there is no parameter for NUTS-2, because the model is done for only one province.

For the variables "Fiscal income", "Non-work income", "Age" and "Sex", parameters from synthetic or from regression model have almost always the same sign, only the amplitude can differ. Parameters for the household size are unstable between the two models.

[7] Because of technical issues, we cannot have convergence with PROC GLIMMIX with random effects. Temporarily, we considers these effects as fixed. We can do that without any other technical issues, because that add a weak amount of degrees of freedom in view of the sample size.

Table 8 : Parameters estimation for AROP, SILC 2016.

NUTS-2	Parameter	Synthetic model	Regression model	EBLUP estimation
BE10 (Brussels)	Intercept	2,78187675	1,68981867	2,7371968
	NUTS-2	-0,02926853		-0,02926853
	Fiscal income	-0,00024548	-0,00020532	-0,00024384
	Non-work income	1,6126E-05	1,701E-05	1,6162E-05
	Age	-0,00986478	-0,00569833	-0,00969432
	Age (square)	4,5634E-05	2,3995E-05	4,4749E-05
	Sex	0,05173489	0,04707002	0,05154403
	Household size (2)	-0,24300702	0,28440251	-0,22142884
	Household size (3)	0,11318914	0,74691158	0,13911696
	Household size (4)	-0,41008799	0,40843033	-0,37659952
BE21 (Antwerpen)	Intercept	2,78187675	2,78014187	2,7818241
	NUTS-2	-0,33175585		-0,33175585
	Fiscal income	-0,00024548	-0,00023633	-0,00024521
	Non-work income	1,6126E-05	-1,5275E-05	1,5173E-05
	Age	-0,00986478	-0,007256	-0,00978562
	Age (square)	4,5634E-05	6,6703E-05	4,6273E-05
	Sex	0,05173489	-0,06961723	0,0480524
	Household size (2)	-0,24300702	-0,33970777	-0,24594146
	Household size (3)	0,11318914	0,09035675	0,11249628
	Household size (4)	-0,41008799	-16,9278274	-0,91132727
BE22 (Limburg)	Intercept	2,78187675	2,09172887	2,77093638
	NUTS-2	-0,26972891		-0,26972891
	Fiscal income	-0,00024548	-0,00032472	-0,00024674
	Non-work income	1,6126E-05	9,4176E-05	1,7363E-05
	Age	-0,00986478	0,00843668	-0,00957467
	Age (square)	4,5634E-05	-0,00012443	4,2938E-05
	Sex	0,05173489	0,1466772	0,05323993
	Household size (2)	-0,24300702	-0,83932941	-0,25246005
	Household size (3)	0,11318914	0,11746228	0,11325688
	Household size (4)	-0,41008799	1,89195035	-0,37359559
BE23 (Oost Vlaanderen)	Intercept	2,78187675	1,55618356	2,74470325
	NUTS-2	-0,53133349		-0,53133349
	Fiscal income	-0,00024548	-0,00024577	-0,00024549
	Non-work income	1,6126E-05	1,7179E-05	1,6158E-05
	Age	-0,00986478	-0,00344762	-0,00967016
	Age (square)	4,5634E-05	8,1472E-05	4,6721E-05
	Sex	0,05173489	-0,057159	0,04843229
	Household size (2)	-0,24300702	0,3243541	-0,22579978
	Household size (3)	0,11318914	0,74736787	0,13242286
	Household size (4)	-0,41008799	-14,8894358	-0,84922566
BE24 (Vlaams Brabant)	Intercept	2,78187675	0,2095073	2,72622197
	NUTS-2	-0,83987532		-0,83987532
	Fiscal income	-0,00024548	-0,00016605	-0,00024377
	Non-work income	1,6126E-05	1,3246E-05	1,6064E-05
	Age	-0,00986478	-0,00275059	-0,00971086
	Age (square)	4,5634E-05	-5,909E-05	4,3368E-05
	Sex	0,05173489	0,08190291	0,05238759
	Household size (2)	-0,24300702	0,28343935	-0,23161703
	Household size (3)	0,11318914	1,02399436	0,13289496
	Household size (4)	-0,41008799	0,17010418	-0,39753518
BE25 (West Vlaanderen)	Intercept	2,78187675	1,06306875	2,73419056
	NUTS-2	-0,52588692		-0,52588692
	Fiscal income	-0,00024548	-0,00014894	-0,00024281

	Non-work income	1,6126E-05	-1,2156E-05	1,5341E-05
	Age	-0,00986478	-0,00712913	-0,00978889
	Age (square)	4,5634E-05	5,4895E-05	4,5891E-05
	Sex	0,05173489	0,00517747	0,05044321
	Household size (2)	-0,24300702	0,01051706	-0,23597331
	Household size (3)	0,11318914	0,03064811	0,11089914
	Household size (4)	-0,41008799	-18,5763496	-0,91408828
BE31 (Brabant Wallon)	Intercept	2,78187675	3,61248712	2,78822525
	NUTS-2	-0,46444428		-0,46444428
	Fiscal income	-0,00024548	-0,00023856	-0,00024543
	Non-work income	1,6126E-05	-1,2984E-05	1,5904E-05
	Age	-0,00986478	-0,02899092	-0,01001097
	Age (square)	4,5634E-05	5,3658E-05	4,5695E-05
	Sex	0,05173489	0,21585685	0,0529893
	Household size (2)	-0,24300702	-0,70308981	-0,24652351
	Household size (3)	0,11318914	-1,84608304	0,09821408
	Household size (4)	-0,41008799	-16,8698158	-0,53589255
BE32 (Hainaut)	Intercept	2,78187675	3,59474272	2,80680543
	NUTS-2	0,0097246		0,0097246
	Fiscal income	-0,00024548	-0,00029944	-0,00024714
	Non-work income	1,6126E-05	4,5364E-05	1,7023E-05
	Age	-0,00986478	0,01098661	-0,00922532
	Age (square)	4,5634E-05	-0,00032134	3,438E-05
	Sex	0,05173489	0,09277404	0,05299346
	Household size (2)	-0,24300702	-0,76774349	-0,25909945
	Household size (3)	0,11318914	-0,8859979	0,08254643
	Household size (4)	-0,41008799	1,17336466	-0,36152724
BE33 (Liège)	Intercept	2,78187675	3,33058632	2,79397569
	NUTS-2	-0,09556618		-0,09556618
	Fiscal income	-0,00024548	-0,00031743	-0,00024707
	Non-work income	1,6126E-05	6,7871E-06	1,592E-05
	Age	-0,00986478	-0,01430695	-0,00996273
	Age (square)	4,5634E-05	0,00012866	4,7465E-05
	Sex	0,05173489	0,31066126	0,05744416
	Household size (2)	-0,24300702	0,2929778	-0,23118866
	Household size (3)	0,11318914	-0,3059554	0,10394708
	Household size (4)	-0,41008799	-14,5573826	-0,72203307
BE34 (Luxembourg)	Intercept	2,78187675	1,64534665	2,77553674
	NUTS-2	-0,77515929		-0,77515929
	Fiscal income	-0,00024548	-0,00011426	-0,00024475
	Non-work income	1,6126E-05	2,6041E-05	1,6181E-05
	Age	-0,00986478	-0,0402953	-0,01003454
	Age (square)	4,5634E-05	0,0002875	4,6983E-05
	Sex	0,05173489	-0,09705192	0,0509049
	Household size (2)	-0,24300702	-1,23388178	-0,24853451
	Household size (3)	0,11318914	-0,85782475	0,10777244
	Household size (4)	-0,41008799	0	-0,40780036
BE35 (Namur)	Intercept	2,78187675	2,42403531	2,77928796
	NUTS-2	0		0
	Fiscal income	-0,00024548	-0,00019066	-0,00024509
	Non-work income	1,6126E-05	4,1869E-05	1,6312E-05
	Age	-0,00986478	-0,00473139	-0,00982765
	Age (square)	4,5634E-05	-6,0349E-05	4,4867E-05
	Sex	0,05173489	0,064099	0,05182434
	Household size (2)	-0,24300702	-0,84852827	-0,24738764
	Household size (3)	0,11318914	-2,37422751	0,09519402
	Household size (4)	-0,41008799	0,26754308	-0,40518569

4.2.3 Application of the model to the entire population

For each poverty indicators, we have now a predictive model based on administrative data available for the entire population. So we apply the model for the year N and obtain, for each individual in Belgium, a probability to be poor according to one of the poverty indicators.

Next, we generate a random number and, by comparison between this random number and the probability of being poor, we obtain an estimation of poverty for each individual and for each poverty indicator, according to our model. Then we aggregate results by NUTS-2 and obtain the poverty rate by NUTS-2.

We need to emphasize that individual prediction of poverty have no sense. Only the aggregation into NUTS-2 could be analyzed, because we proved before that this model is relevant.

4.2.4 Regional correction of the indicators

The NUTS-2 indicators of poverty are modeled but their aggregation at NUTS-1 or at national level are not equal to the indicator obtained by the direct estimation on the sample. Despite that, Statistics Belgium need to provide coherent results across NUTS, so we have to revise the indicators based on regional results. Table 9 highlights these differences for the estimation of AROP for SILC 2016.

Table 9 : Regional discordance between direct and EBLUP estimation (AROP, SILC 2016)

	N (sample)	N (pop)	Direct estimation	EBLUP
BE10 (Brussels)	2365	1.117.922	31,1%	35,1%
AVERAGE BRUSSELS			31,1%	35,1%
BE21 (Antwerpen)	1735	1.788.021	11,7%	13,5%
BE22 (Limburg)	893	851.433	10,6%	12,4%
BE23 (Oost Vlaanderen)	1734	1.461.042	9,0%	10,1%
BE24 (Vlaams Brabant)	1226	1.096.081	8,9%	8,7%
BE25 (West Vlaanderen)	1582	1.162.100	12,0%	9,5%
AVERAGE FLANDERS			10,5%	10,9%
BE31 (Brabant Wallon)	427	388.282	10,9%	12,8%
BE32 (Hainaut)	1754	1.307.657	22,9%	23,5%
BE33 (Liège)	1250	1.079.496	18,9%	21,4%
BE34 (Luxembourg)	311	274.911	11,2%	11,9%
BE35 (Namur)	404	481.170	22,4%	19,2%
AVERAGE WALLONIA			19,3%	20,2%

The final weight (after the calibration), for an individual *i*, a NUTS-2 *j* and a NUTS-1 *k* (each NUTS-2 can belong to only one NUTS-1), is $w_{i,j,k}$. The weight of the NUTS-2 (i.e. the number of individuals into it) is $w_{j,k} = \sum_i w_{i,j,k}$ .Similarly, the weight of a NUTS-1 is $w_k = \sum_j w_{j,k}$ .These two last weights have to remain equal in order to keep a concordance between NUTS-1 and NUTS-2 results.

According to the previous section, $\bar{Y}_{j,k}^{E}$ is the NUTS-2 EBLUP estimator before the regional correction, and $\bar{Y}_{j,k}^{E\prime}$ after the regional correction. $\bar{Y}_{k}$ is the NUTS-1 direct estimator of the poverty indicator calculated by the survey, and $\bar{Y}_{k}^{E}$ is the NUTS-1 EBLUP estimator of the poverty indicator.

In order to keep unchanged the rank between NUTS-2 estimators, while ensuring regional concordance of the results, we use a simple rule of three to adjust the estimators :

$$\bar{Y}_{j,k}^{E\prime} = \bar{Y}_{j,k}^{E} \frac{\bar{Y}_{k}}{\bar{Y}_{k}^{E}}$$

The final EBLUP estimators of poverty $\bar{Y}_{j,k}^{E\prime}$ will be close to their initial value, and regional concordance will be guarantee. Table 10 shows how the correction is done for the AROP estimation for SILC 2016. For Flanders, the correction factor is 0,9502. For Wallonia, it is 0,9569. For Brussels, the corrected EBLUP is equal to the direct estimation, because Brussels is both a NUTS-1 and NUTS-2.

Table 10 : Regional correction of EBLUP estimators (AROP, SILC 2016)

	n (sample)	N (pop)	Direct estimation	EBLUP	Corrected EBLUP
BE10 (Brussels)	2.365	1.117.922	31,1%	35,1%	31,1%
AVERAGE BRUSSELS			**31,1%**	**35,1%**	**31,1%**
BE21 (Antwerpen)	1.735	1.788.021	11,7%	13,5%	12,8%
BE22 (Limburg)	893	851.433	10,6%	12,4%	11,8%
BE23 (Oost Vlaanderen)	1734	1.461.042	9,0%	10,1%	9,6%
BE24 (Vlaams Brabant)	1226	1.096.081	8,9%	8,7%	8,2%
BE25 (West Vlaanderen)	1582	1.162.100	12,0%	9,5%	9,0%
AVERAGE FLANDERS			**10,5%**	**11.0%**	**10,5%**
BE31 (Brabant Wallon)	427	388.282	10,9%	12,8%	12,2%
BE32 (Hainaut)	1754	1.307.657	22,9%	23,5%	22,5%
BE33 (Liège)	1250	1.079.496	18,9%	21,4%	20,5%
BE34 (Luxembourg)	311	274.911	11,2%	11,9%	11,4%
BE35 (Namur)	404	481.170	22,4%	19,2%	18,4%
AVERAGE WALLONIA			**19,3%**	**20,2%**	**19,3%**

4.2.5 Adjustments on individual data

As required by Eurostat, we will deliver an individual database allowing to calculate our final EBLUP estimator. This is only a technical constraint, and the indicators of poverty for each individual do not have to be relevant for other purpose. We solve this technical constraint as following: based on the sample database, we redress the individual weight in order to obtain the final estimator of the poverty indicators. We proceed with the macro CALMAR, with 4 margins by NUTS-2: proportion of AROP, LWI,

SMD and AROPE. For each NUTS-2, we redress individual weight only to get these 4 NUTS-2 estimators of poverty.

With this new weighing set, it is only possible to calculate the EBLUP estimator of each poverty indicators at NUTS-2 level. We must stress the fact that these weights are only relevant to calculate EBLUP provided by our model. It is not possible to estimate poverty for a new sample size, or for a sub-population inside a NUTS-2.

4.3 First results: NUTS-2 by SAE for SILC 2011 to 2016

We apply here the methodology explained in the previous sub-section to obtain EBLUP estimators at NUTS-2 level for SILC 2011 to 2016. To provide our model, here is the data we used:

- EU-SILC survey from 2009 to 2016
- Explanatory variables: fiscal income, fiscal LWI, age, sex, household size
- Explained variables: AROP, LWI, SMD, AROPE
- For each year N, we use the data from N-2 to N

We focus here mostly on the most recent results, namely those of 2016. However, in order to assess the quality of our methodology, we provide results from 2011 to 2015 too. So we can highlight the instability of the direct estimators and the quite stability of our EBLUP estimators. As we explained before, EBLUP estimators are mostly based on the synthetics estimators (see table 7).

4.3.1 AROP and AROPE

Table 11 (next page) summarizes AROP estimator at NUTS-2 level for 2016. Differences between direct and EBLUP estimators are weak (between 0.2 and 3 pp.). The corrected EBLUP estimator is near the direct estimator too. While EBLUP is mostly based on synthetic estimator, the model provides concordant results with the direct estimator provided by the sample. By assuming that the sample is not biased, this is another proof the quality of our model.

The bigger gap between direct and EBLUP estimator concerns the province of Namur (22.4 % for the direct estimation versus 19.2 % for the EBLUP estimation, see table 11). For 2016, the means of fiscal income is 20157€ into the sample, versus 21207 € into the population (see table 1). The fiscal income is negatively correlated with AROP, so we could except that an estimator based on population data will be weaker than an estimator based on sample data. The other auxiliary variables (non-work income, age, gender, household size) are mostly identical between the sample and the population.

Table 11 : AROP direct and EBLUP estimation by NUTS-2, SILC 2016

	Direct estimation	EBLUP	Corrected EBLUP
BE10 (Brussels)	31,1%	35,1%	31,1%
BE21 (Antwerpen)	11,7%	13,5%	12,8%

BE22 (Limburg)	10,6%	12,4%	11,8%
BE23 (Oost Vlaanderen)	9,0%	10,2%	9,6%
BE24 (Vlaams Brabant)	8,9%	8,7%	8,2%
BE25 (West Vlaanderen)	12,0%	9,5%	9,0%
BE31 (Brabant Wallon)	10,9%	12,8%	12,2%
BE32 (Hainaut)	22,9%	23,5%	22,5%
BE33 (Liège)	18,9%	21,4%	20,5%
BE34 (Luxembourg)	11,2%	11,9%	11,4%
BE35 (Namur)	22,4%	19,2%	18,4%

Brussels is a special case. Indeed, it is at the same time a province (NUTS-2) and a region (NUTS-1). So, the sample of Brussels is used in the model, but the regional correction implies that the EBLUP estimator is identical with the direct estimation. For this region, we use another plan to improve the precision of the direct estimation, by changing the sampling design.

Even if EBLUP estimators are near to direct estimators for 2016, we can notice that the AROP indicator is more stable over the years. In the next page, you can see an illustration for the AROP indicator, with direct and with EBLUP estimator (figure 6 and 7).

Each year is separately analyzed, and the result is not like a moving average. And even if results are more stable, the rank of a NUTS-2 can change year to year. For example, between 2015 and 2016, AROP for Brabant-Wallon became higher than AROP for Antwerpen, these two indicators remain very close however (the difference is less than 1 pp.).

We can group the NUTS-2 into four clusters:

- Brussels (BE10), with the higher rate of AROP. As we explained before, Brussels is at the same time a NUTS-1 and a NUTS-2. Thus, the direct and corrected EBLUP estimation are similar. The AROP rate is quite stable ;
- Namur (BE35), Hainaut (BE32) and Liège (BE33), with an AROP rate around 20%. In the direct estimation, these provinces are very fluctuant but they remain together (especially in 2011 and in 2016). In the EBLUP estimation, these provinces remain together all the time ;
- Luxembourg (BE34), Brabant Wallon (BE31), Antwerpen (BE21) and Limburg (BE22), with an AROP rate around 12%. The EBLUP estimation for these provinces is very stable over the time ;
- Oost Vlaanderen (BE23), West Vlaanderen (BE25), Vlaams Brabant (BE24), with an AROP rate around 9%. The EBLUP estimation for these provinces is very stable over the time too.

Figure 6 : AROP direct estimation by NUTS-2, SILC 2011 - 2016

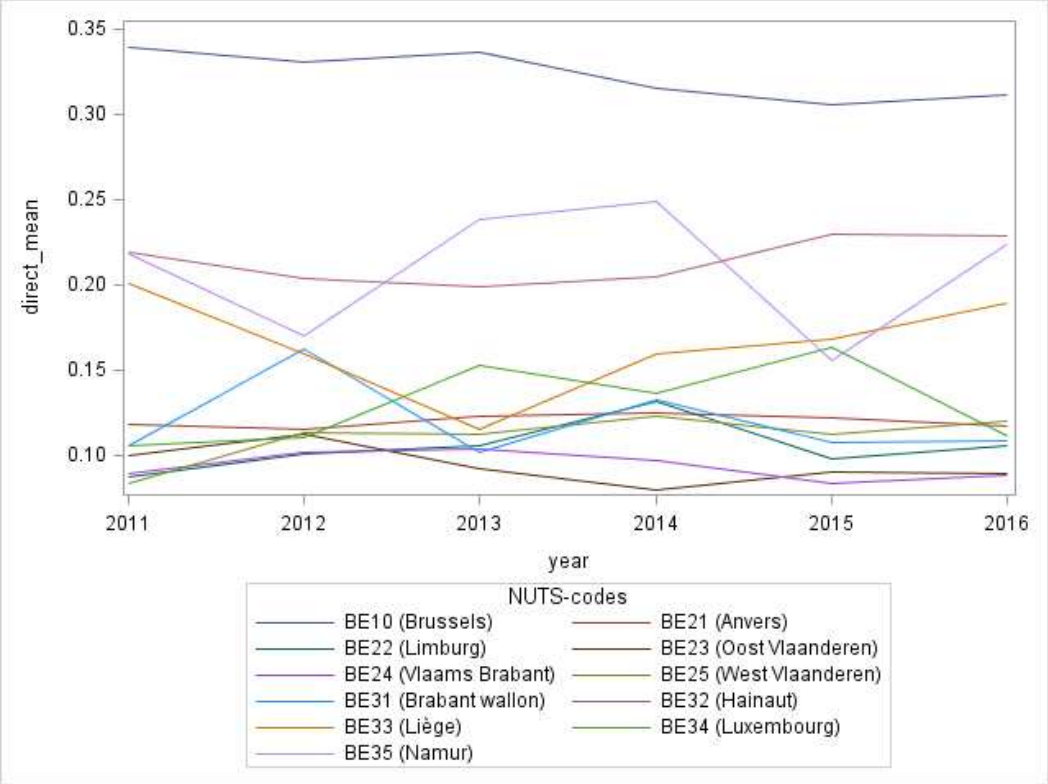
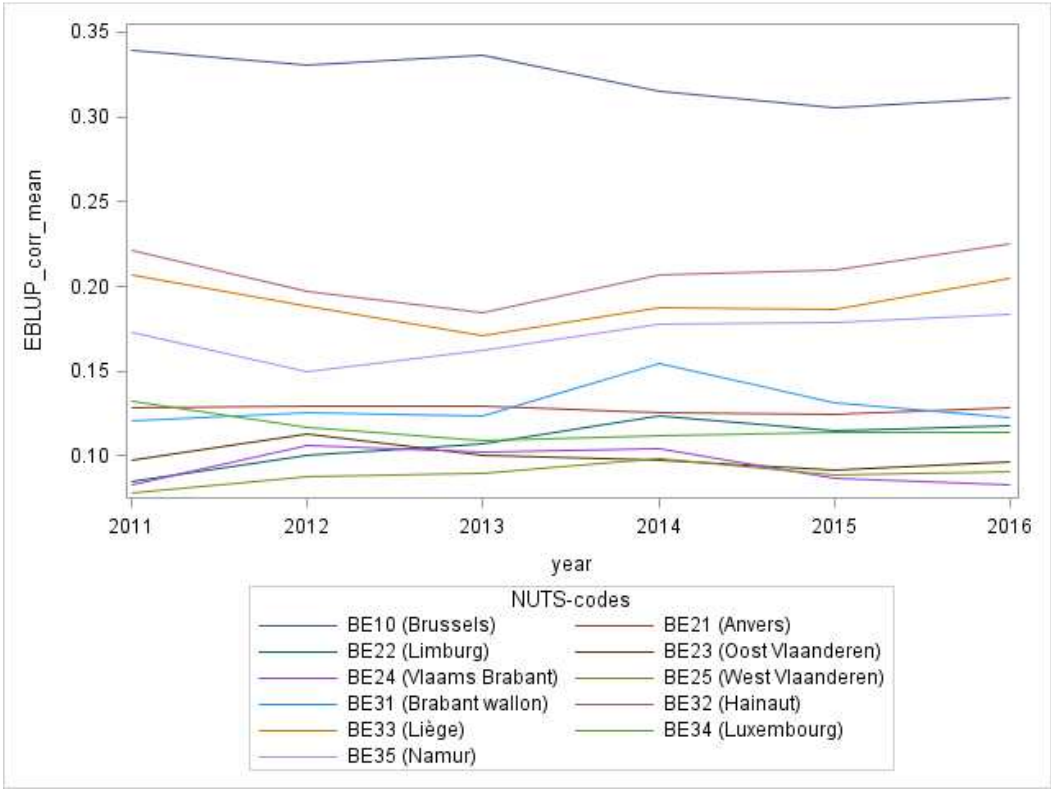


Figure 7 : AROP EBLUP estimation by NUTS-2, SILC 2011 - 2016



The analysis for the AROPE indicator is similar. For 2016, the direct and the EBLUP estimation of AROPE is quite the same, except for few NUTS-2 (see table 14). The regional correction is weak and the indicator is stable over the time. The gap between the two estimators can be explained by the difference of the auxiliary variables between the sample and the entire population. Figure 8 and 9 illustrate the effect of the EBLUP estimation from 2011 to 2016 for AROPE. The four clusters defined for AROP are quite similar for AROPE.

Table 12 : AROPE direct and EBLUP estimation by NUTS-2, SILC 2016

	Direct estimation	EBLUP	Corrected EBLUP
BE10 (Brussels)	38.30%	42.86%	38.30%
BE21 (Antwerpen)	16.19%	18.19%	17.06%
BE22 (Limburg)	15.32%	18.21%	17.08%
BE23 (Oost Vlaanderen)	12.94%	14.55%	13.65%
BE24 (Vlaams Brabant)	12.74%	12.68%	11.89%
BE25 (West Vlaanderen)	14.94%	13.00%	12.20%
BE31 (Brabant Wallon)	19.22%	19.06%	18.02%
BE32 (Hainaut)	29.42%	32.08%	30.34%
BE33 (Liège)	26.66%	29.34%	27.75%
BE34 (Luxembourg)	16.47%	16.30%	15.42%
BE35 (Namur)	27.54%	25.51%	24.12%

Figure 8 : AROPE direct estimation by NUTS-2, SILC 2011 - 2016

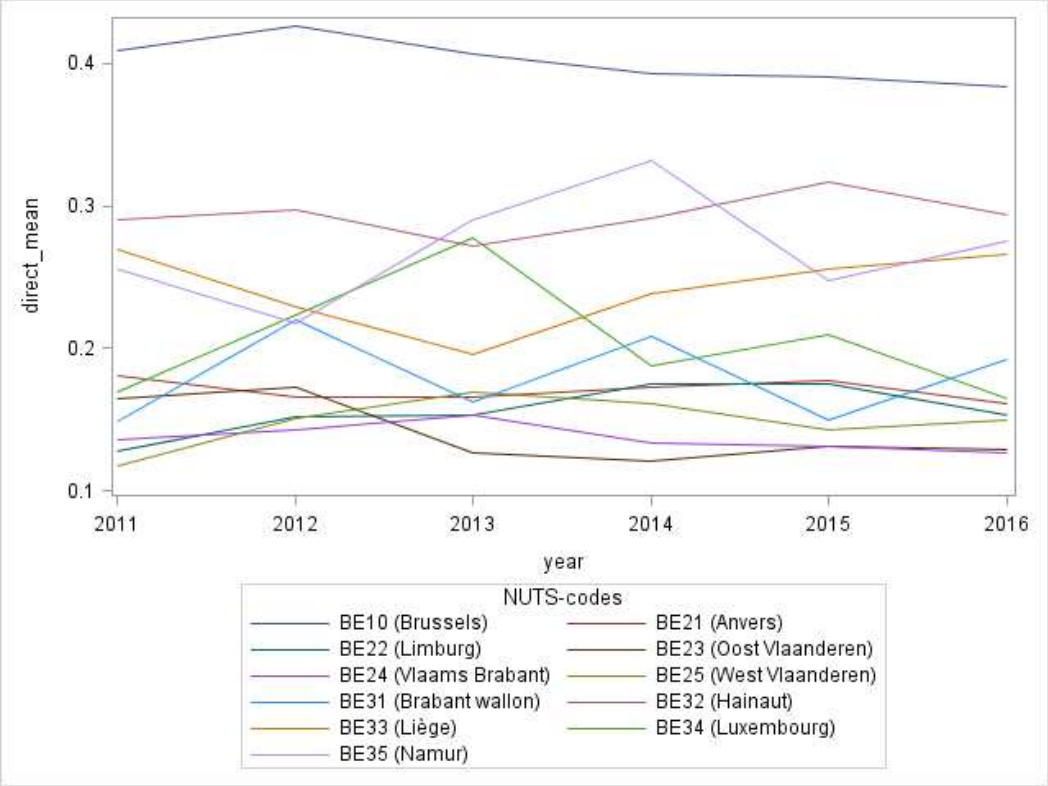
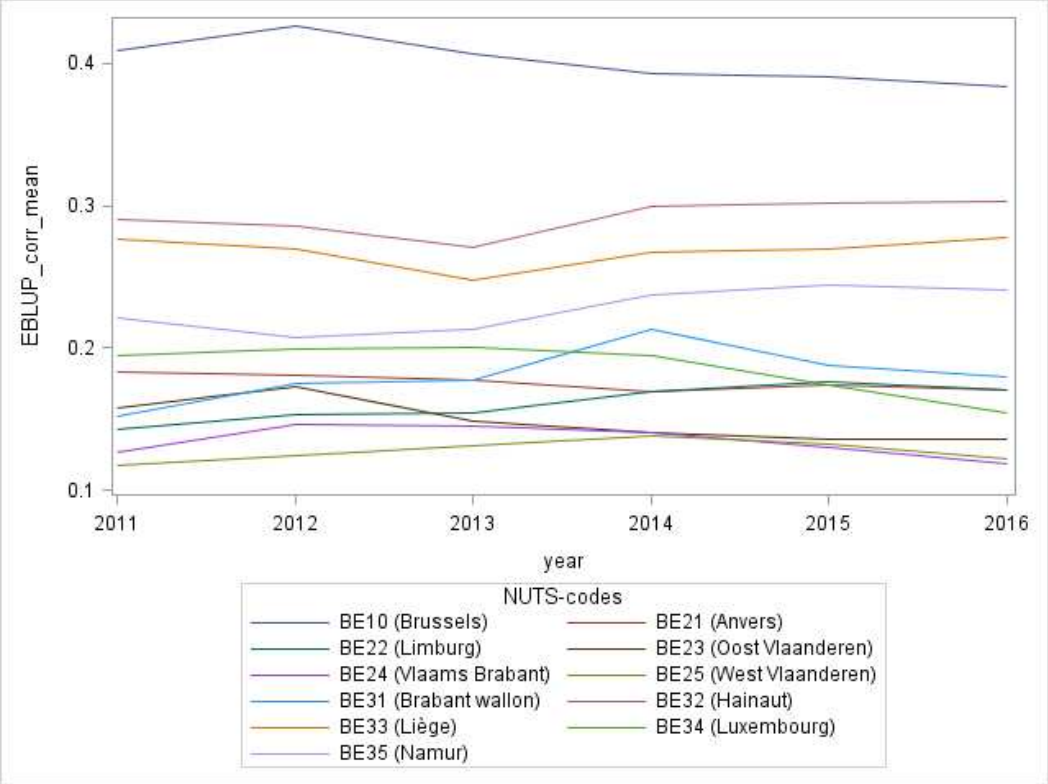


Figure 9 : AROPE EBLUP estimation by NUTS-2, SILC 2011 - 2016



4.3.2 LWI

The LWI direct estimator is very instable at NUTS-2 level. The figure 7 shows the evolution of this estimator between 2011 and 2016. Some differences are more than 10 pp. from one year to another, which can be caused by the small sample size (the sample for the LWI indicator is smaller than for the others indicators, because age of individuals must be between 16 and 60). So, EBLUP estimator is more important here to produce suitable results for the LWI indicator.

For the year 2016, EBLUP estimators are very different from direct estimators based on the sample (table 12). So, the regional correction is stronger for LWI than for AROP or AROPE. We can see the same fact for each year. Thus, EBLUP estimators for LWI are more stable between 2011 and 2016 (figure 8). These results do not show the same cluster as for AROP or AROPE. The LWI poverty indicator is independent from AROP poverty indicator, so provinces do not have to display the same profile.

Figure 10 : LWI direct estimation by NUTS-2, SILC 2011 - 2016

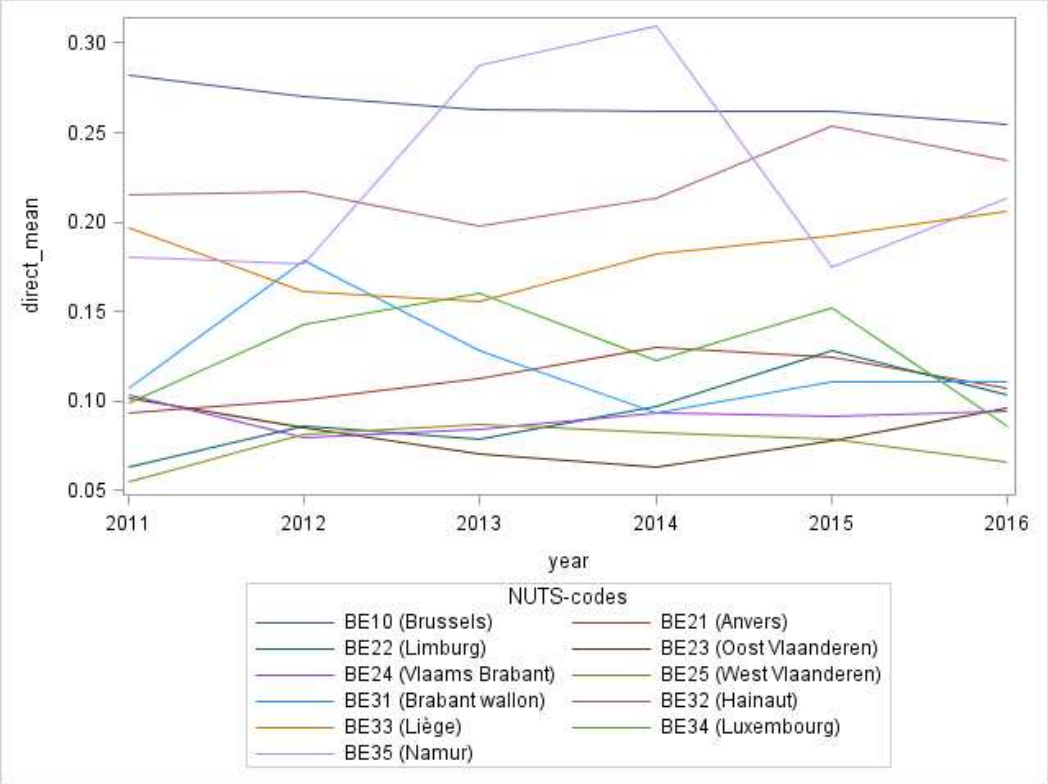


Figure 11 : LWI EBLUP estimation by NUTS-2, SILC 2011 - 2016

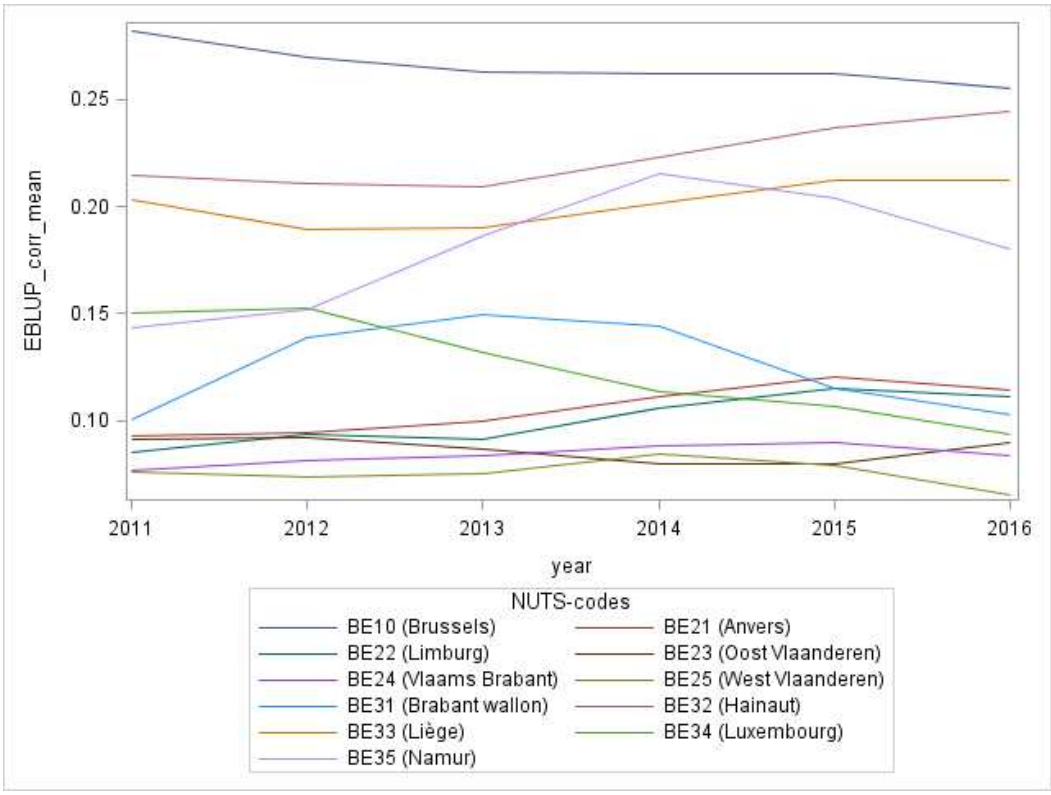


Table 13 : LWI direct and EBLUP estimation by NUTS-2, SILC 2016

	Direct estimation	EBLUP	Corrected EBLUP
BE10 (Brussels)	25.48%	27.87%	25.48%
BE21 (Antwerpen)	10.73%	13.06%	11.48%
BE22 (Limburg)	10.35%	12.64%	11.10%
BE23 (Oost Vlaanderen)	9.65%	10.27%	9.02%
BE24 (Vlaams Brabant)	9.47%	9.57%	8.41%
BE25 (West Vlaanderen)	6.56%	7.43%	6.53%
BE31 (Brabant Wallon)	11.05%	10.62%	10.32%
BE32 (Hainaut)	23.43%	25.18%	24.47%
BE33 (Liège)	20.64%	21.84%	21.22%
BE34 (Luxembourg)	8.57%	9.69%	9.41%
BE35 (Namur)	21.35%	18.54%	18.01%

4.3.3 SMD

For the SMD indicator, we saw in section II that the auxiliary variables do not explain this entire poverty indicator. Thus, we know that our estimation of SMD by NUTS-2 could be problematic. For 2016, we notice that the difference between direct and EBLUP estimation is very weak, except for Namur (once again, this province has a small sample size and some high differences occur between the sample and the population). We also notice that EBLUP estimator for SMD is more sensitive to the direct estimator. Comparing the figure 10 and 11, the variation year-to-year are reported from direct to EBLUP estimator, so the EBLUP estimator have some incredible variations. This problem will be corrected in the future by introducing new auxiliary variables more linked to SMD. From now, all the possible variables we test are already correlated with the fiscal income or they concern a too little cluster (pension of invalidity for example). In the future, we will access to the rental status by administrative data for all the population. We hope enhance the model with this variable.

Table 14 : SMD direct and EBLUP estimation by NUTS-2, SILC 2016

	Direct estimation	EBLUP	Corrected EBLUP
BE10 (Brussels)	13.43%	16.55%	13.43%
BE21 (Antwerpen)	4.09%	4.44%	3.88%
BE22 (Limburg)	3.33%	3.47%	3.03%
BE23 (Oost Vlaanderen)	3.20%	3.65%	3.18%
BE24 (Vlaams Brabant)	1.54%	1.66%	1.45%
BE25 (West Vlaanderen)	1.27%	2.21%	1.93%
BE31 (Brabant Wallon)	8.28%	8.32%	6.96%
BE32 (Hainaut)	11.38%	11.29%	9.43%
BE33 (Liège)	6.01%	8.59%	7.17%
BE34 (Luxembourg)	1.73%	2.19%	1.83%
BE35 (Namur)	4.95%	10.48%	8.76%

Figure 12 : SMD direct estimation by NUTS-2, SILC 2011 - 2016

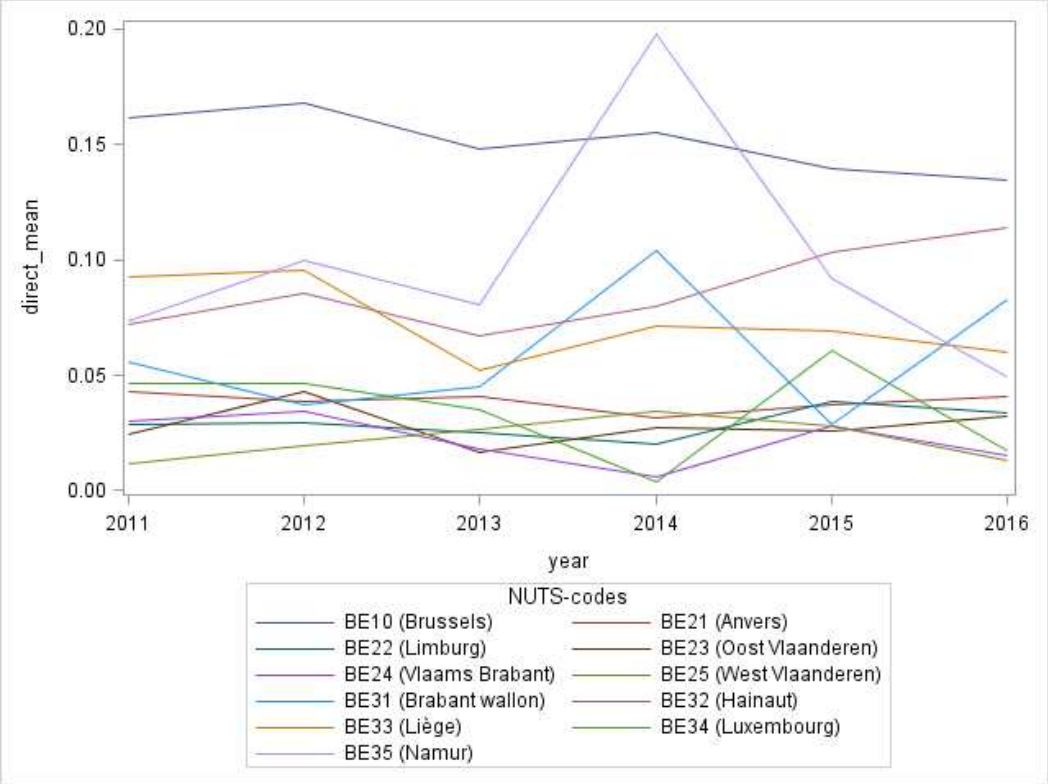
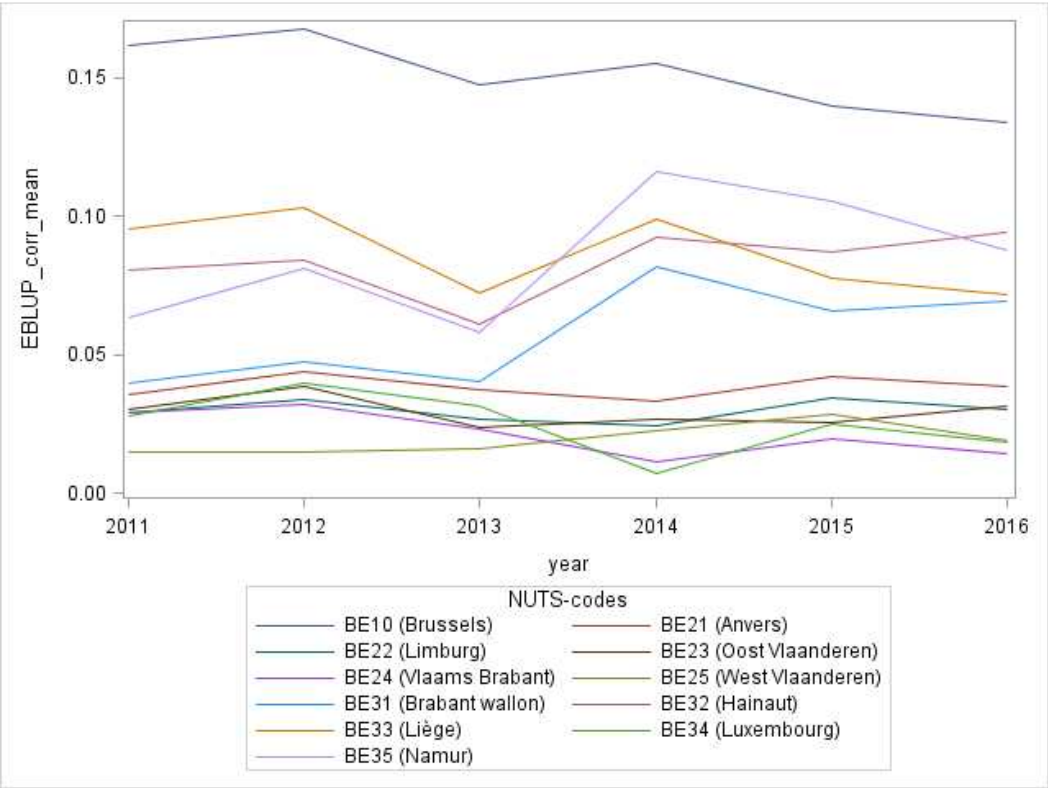


Figure 13 : SMD EBLUP estimation by NUTS-2, SILC 2011 - 2016



5 Conclusion

The main objective was to provide results for all poverty indicators and for all NUTS-2. To achieve this objective, the methodological staff in Statistics Belgium worked from 2015 to August 2018. When this report is written, we can note that this objective is completed: we can now provide EBLUP estimators for AROP, LWI, SMD and AROPE at NUTS-2 level and we can assert our methodology by theoretical proof and empirical evidences.

The first part of the work was to confirm that creating a model for our poverty indicators could be relevant. For that, we assumed some simple assumptions about the link between the poverty indicator and the auxiliary variables available for the entire population. We highlight the fact that results differ by NUTS-2, and that the effects of the NUTS-2 can be altered by the auxiliary variables. These variables provide a great explanatory power, so we could plan a model for the poverty indicators. This work was done between September 2015 and June 2016

The second part of the work was mostly theoretical. We had to learn if we can use mixed model and EBLUP in our situation. Thus, we analyze the formalization of this kind of estimators and we adapt the theory to our specific case. NUTS-2 are seen here as random effect and EBLUP can provide a non-biased estimation of poverty, even with non-linear indicator. This work was done between July 2016 and January 2017.

The last part of the work was about implementation of our theoretical model into our statistical software. We listed all the steps between our initial database and the final result for Eurostat, and then we created macro for each step of this list. We needed to deal with some practical problems with non-linear convergence into SAS. We discussed all of these problems into the report and provide some temporary and final solutions. This work was done between January 2017 and August 2017.

At last, we can provide EBLUP estimators of the four poverty indicators for 2016. Moreover, we can test our model for 6 years and so check the trend of our results. We highlight the quite proximity between direct and EBLUP estimation for AROP and AROPE indicators. Our model do not alter so much the direct estimation, while flatten the trend by NUTS-2. For the LWI indicator, the sample is smaller and the correction has to be stronger. The auxiliary variables are highly correlated with LWI, so we can be trustful into our EBLUP results. For SMD, the model has a lower explanatory power, so EBLUP have to be interpreted with caution.

We still have some work to do during 2017. First we need to validate the estimation of the variance for the EBLUP estimators. We did not report our work in this report because we need some more debates and decisions. We have two ways to provide it: by the model or by a Jackknife estimation based on the sample. Second, we need to improve the model for the SMD indicator. We will obtain in the future some more variables for the entire population and we expect that they will be more correlated with SMD. For example, we will obtain the tenant status for each household in Belgium.

This work is a part of a global reform for SILC to improve the quality of the survey. The sampling plan is currently improved by a new stratification at household level using fiscal data instead of primary sampling units of geographical units for Brussels. Administrative data will be used for the correction of the non-response and for the calibration of the final results at NUTS-1 level. We also plan to shorten the survey and replace some questions about incomes by administrative data. This will help reaching precision requirements at NUTS-1 and at NUTS-2 level.

6 Annexes

Table 15: AROP rate by NUTS-2 and by income quintiles (individual level)

	Belgium	NUTS-2										
		BE10 (Brussels)	BE21 (Antwerpen)	BE22 (Limburg)	BE23 (Flandre Orientale)	BE24 (Brabant Flamand)	BE25 (Flandre Occidentale)	BE31 (Brabant Wallon)	BE32 (Hainaut)	BE33 (Liège)	BE34 (Luxembourg)	BE35 (Namur)
All	15 %	33 %	12 %	10 %	11 %	10 %	11 %	16 %	20 %	16 %	11 %	17 %
Income Quantiles												
1	71 %	78 %	68 %	11 %	65 %	61 %	54 %	77 %	74 %	85 %	0 %	94 %
2	75 %	89 %	36 %	90 %	77 %	74 %	67 %	20 %	79 %	79 %	88 %	82 %
3	66 %	57 %	84 %	63 %	50 %	45 %	77 %	71 %	65 %	68 %	63 %	71 %
4	39 %	59 %	31 %	38 %	20 %	48 %	33 %	50 %	38 %	49 %	30 %	15 %
5	26 %	44 %	23 %	21 %	36 %	42 %	4 %	35 %	21 %	13 %	34 %	16 %
6	15 %	18 %	15 %	12 %	22 %	13 %	12 %	0 %	15 %	14 %	0 %	13 %
7	16 %	24 %	5 %	11 %	17 %	6 %	13 %	48 %	27 %	22 %	17 %	0 %
8	8 %	13 %	8 %	3 %	7 %	4 %	9 %	30 %	6 %	7 %	45 %	0 %
9	3 %	3 %	4 %	0 %	6 %	0 %	7 %	0 %	2 %	0 %	0 %	0 %
10	4 %	7 %	1 %	7 %	5 %	10 %	2 %	12 %	3 %	0 %	4 %	6 %
11	3 %	5 %	11 %	3 %	3 %	0 %	0 %	0 %	1 %	6 %	0 %	0 %
12	3 %	2 %	7 %	0 %	0 %	0 %	2 %	0 %	5 %	1 %	0 %	11 %
13	2 %	2 %	2 %	4 %	3 %	3 %	0 %	0 %	0 %	3 %	0 %	0 %
14	1 %	0 %	1 %	0 %	0 %	0 %	2 %	0 %	3 %	1 %	0 %	0 %
15	1 %	0 %	0 %	4 %	3 %	0 %	1 %	0 %	0 %	2 %	0 %	0 %
16	2 %	0 %	0 %	0 %	0 %	2 %	1 %	34 %	2 %	5 %	0 %	0 %
17	1 %	4 %	2 %	0 %	0 %	0 %	0 %	0 %	5 %	0 %	0 %	6 %
18	1 %	0 %	0 %	0 %	1 %	0 %	1 %	7 %	5 %	0 %	0 %	0 %
19	2 %	0 %	1 %	3 %	2 %	1 %	0 %	0 %	3 %	3 %	0 %	6 %
20	1 %	0 %	0 %	3 %	0 %	0 %	4 %	6 %	2 %	0 %	0 %	0 %
Not available	18 %	17 %	20 %	14 %	16 %	0 %	16 %	14 %	22 %	39 %	8 %	19 %

Note : Some fingers appears to be outliers (0 % of AROP in the first IQ in BE34 ; 34 % of AROP in the 16th IQ in BE31). These results come from the too small sample size in these intersections.

Figure 14: Odds ratios by NUTS-2 for AROP rate (model A, SILC 2016)

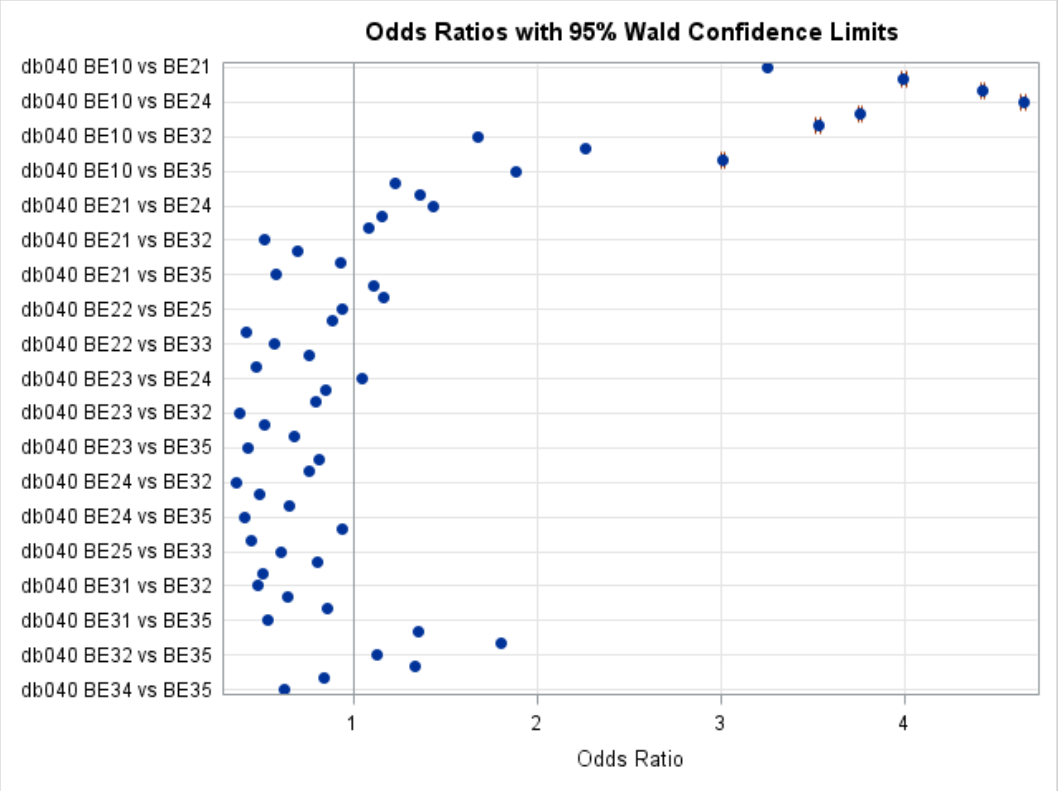


Figure 15 : Odds ratios by NUTS-2 for AROP rate (model C, SILC 2016)

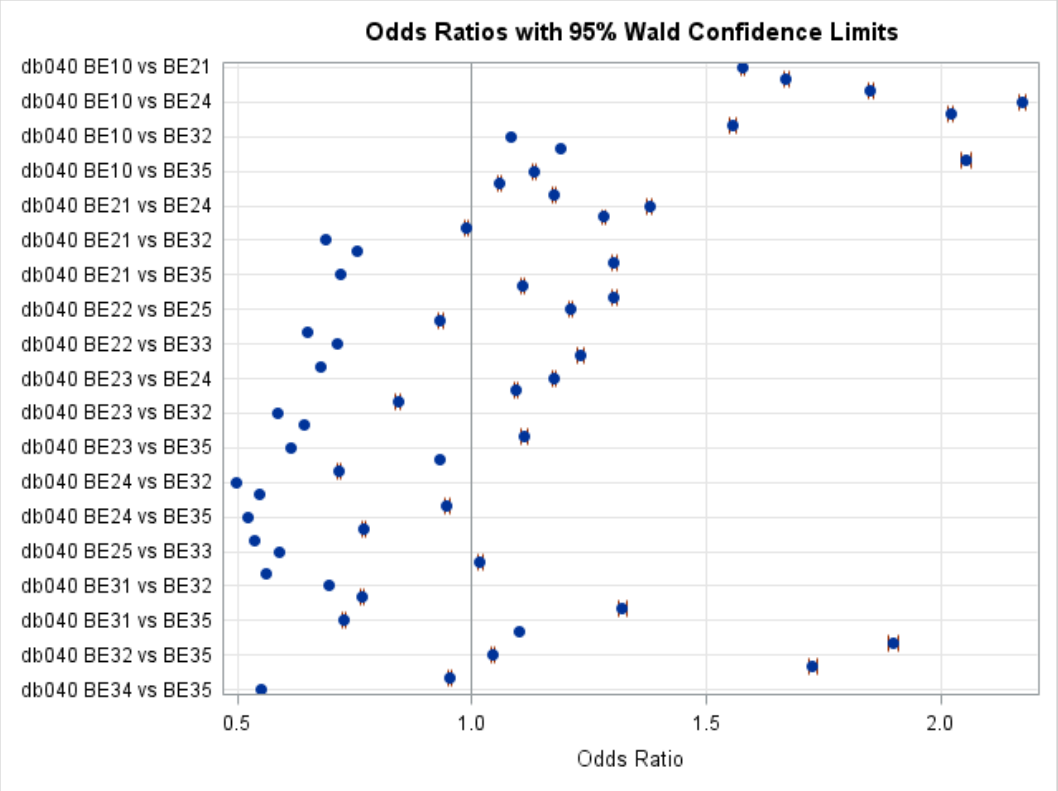


Figure 16: Odds ratios by NUTS-2 for LWI rate (model A, SILC 2016)

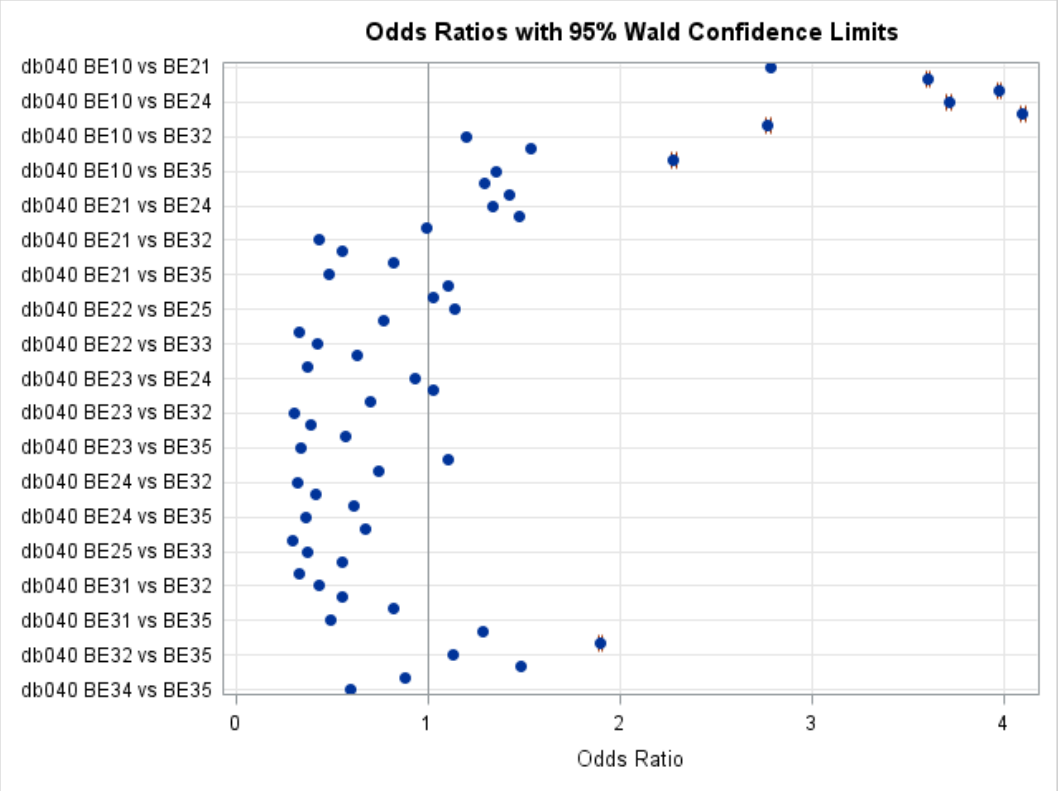


Figure 17: Odds ratios by NUTS-2 for LWI rate (model C, SILC 2016)

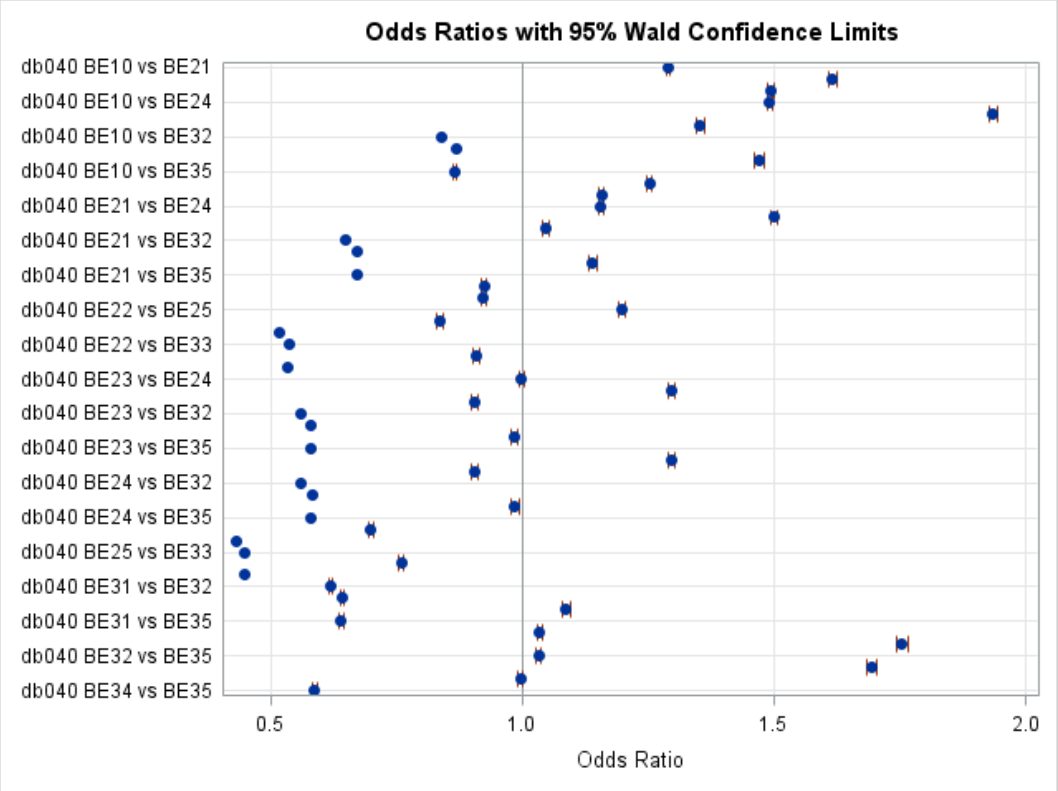


Figure 18: Odds ratios by NUTS-2 for SMD rate (model A, SILC 2016)

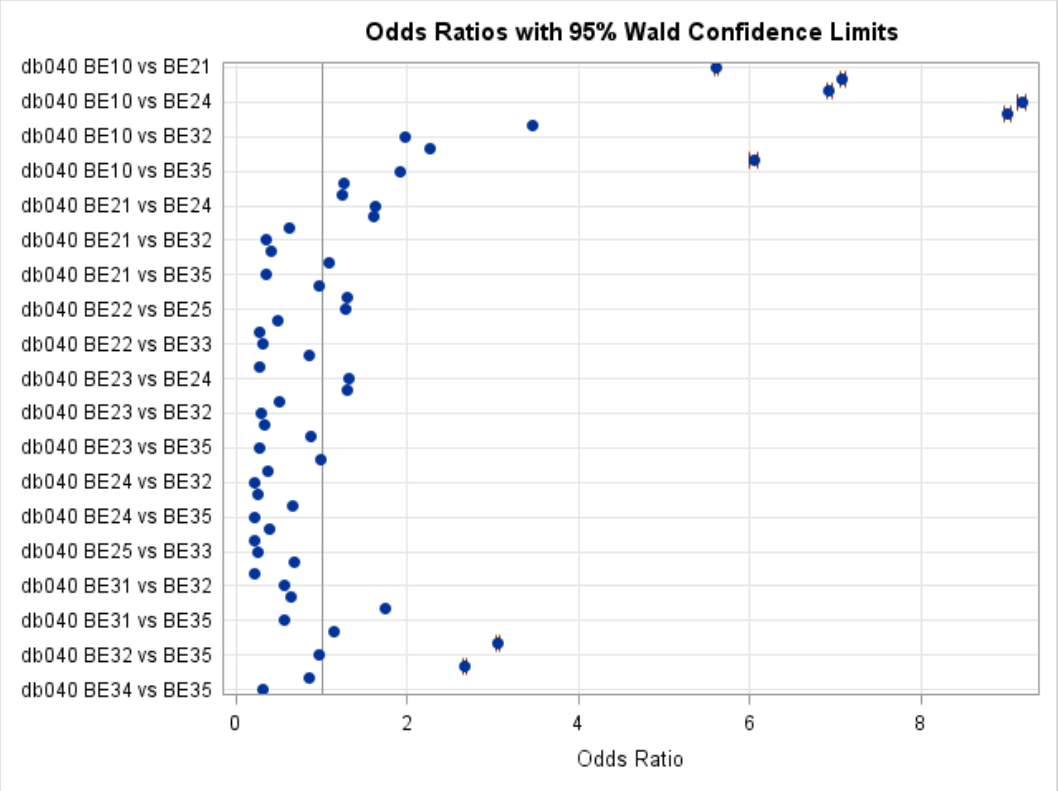


Figure 19 : Odds ratios by NUTS-2 for SMD rate (model C, SILC 2016)

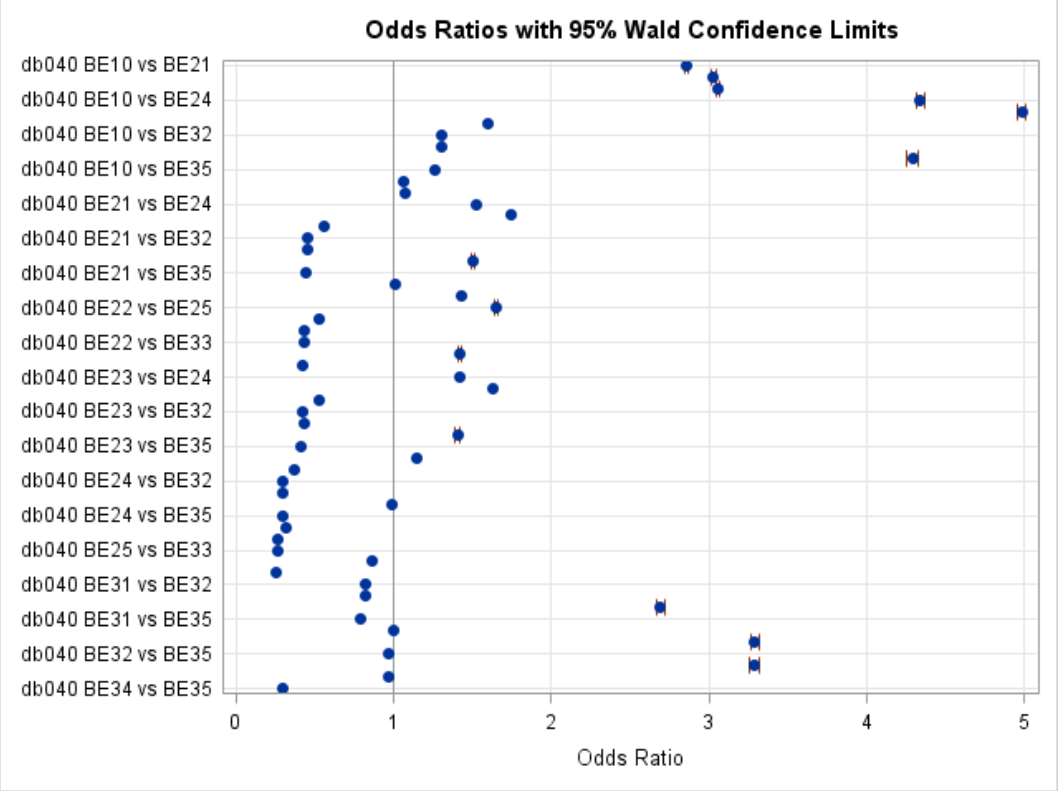


Figure 20: Odds ratios by NUTS-2 for AROPE rate (model A, SILC 2016)

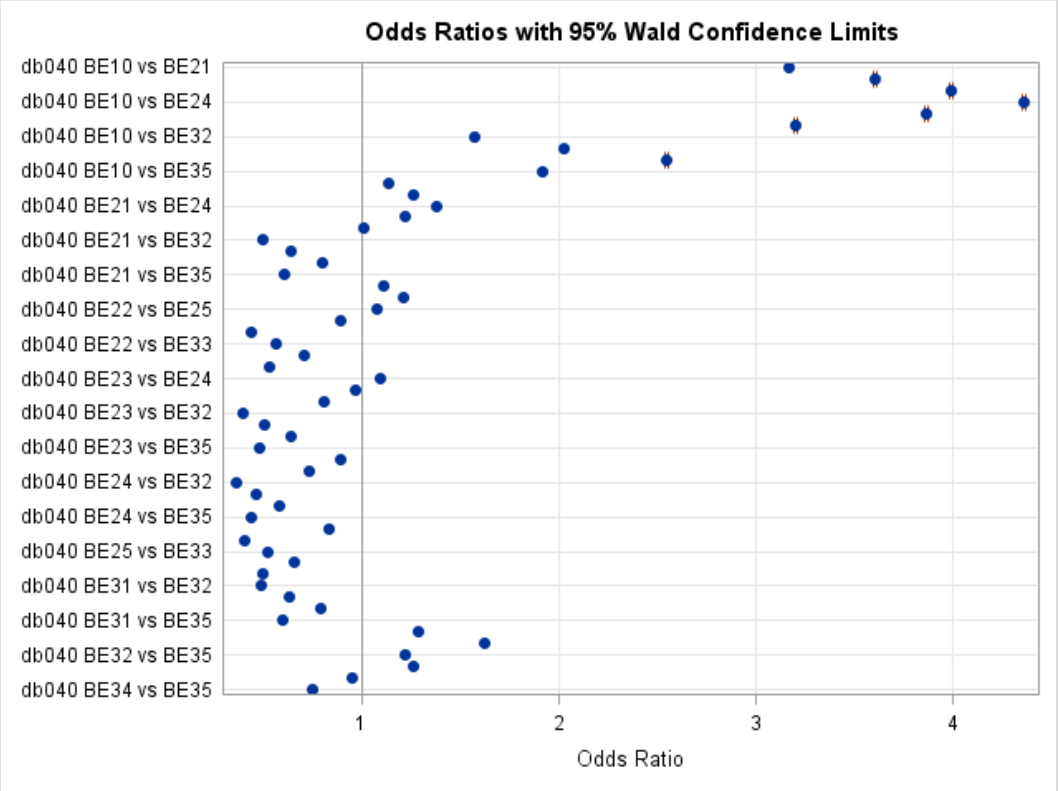


Figure 21 : Odds ratios by NUTS-2 for AROPE rate (model C, SILC 2016)

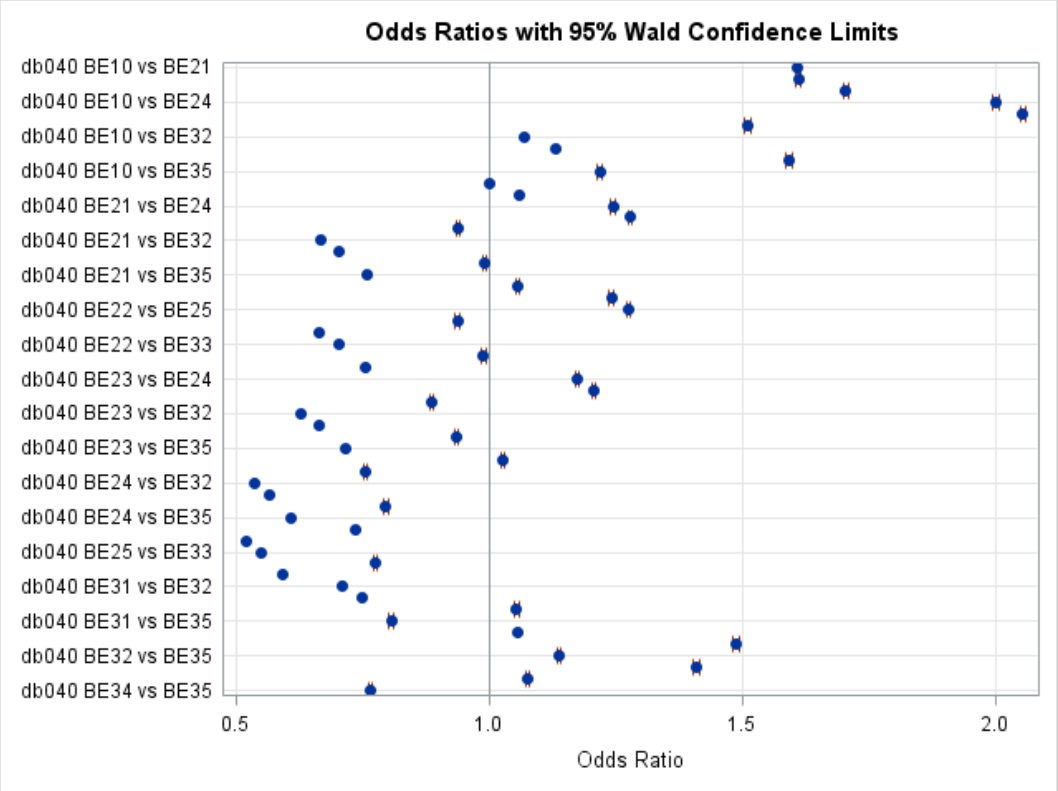


Figure 22: Explanatory power of auxiliary variables for AROP rate (SILC 2016)

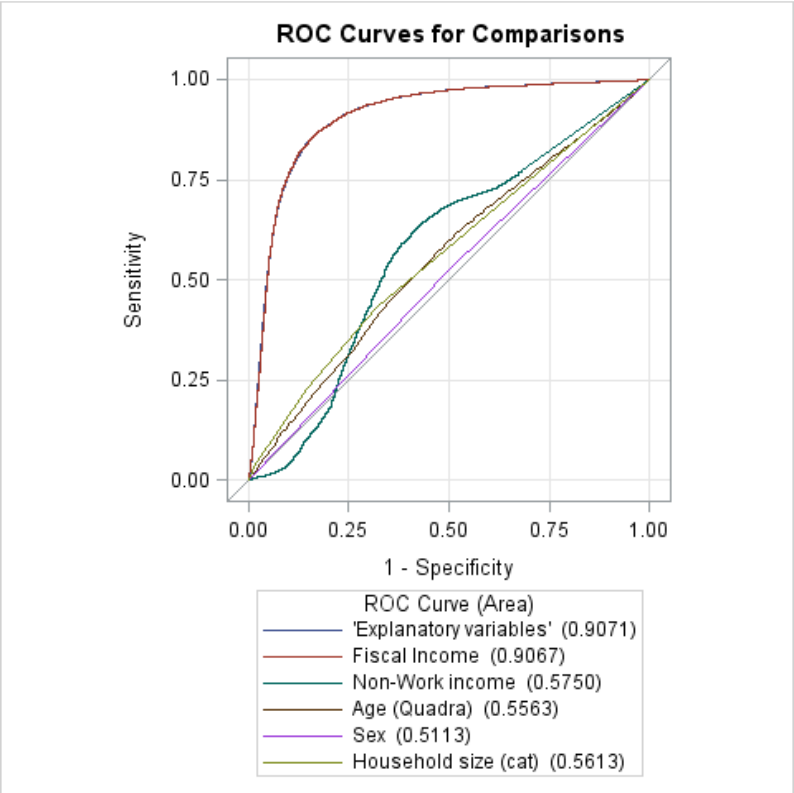


Figure 23: Explanatory power of auxiliary variables for LWI rate (SILC 2016)

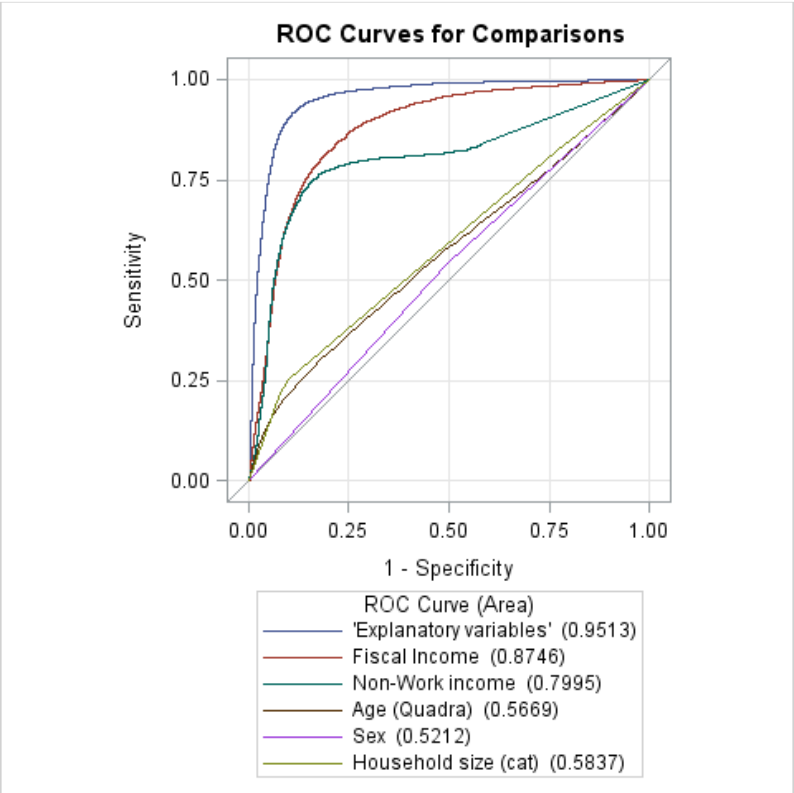


Figure 24: Explanatory power of auxiliary variables for SMD rate (SILC 2016)

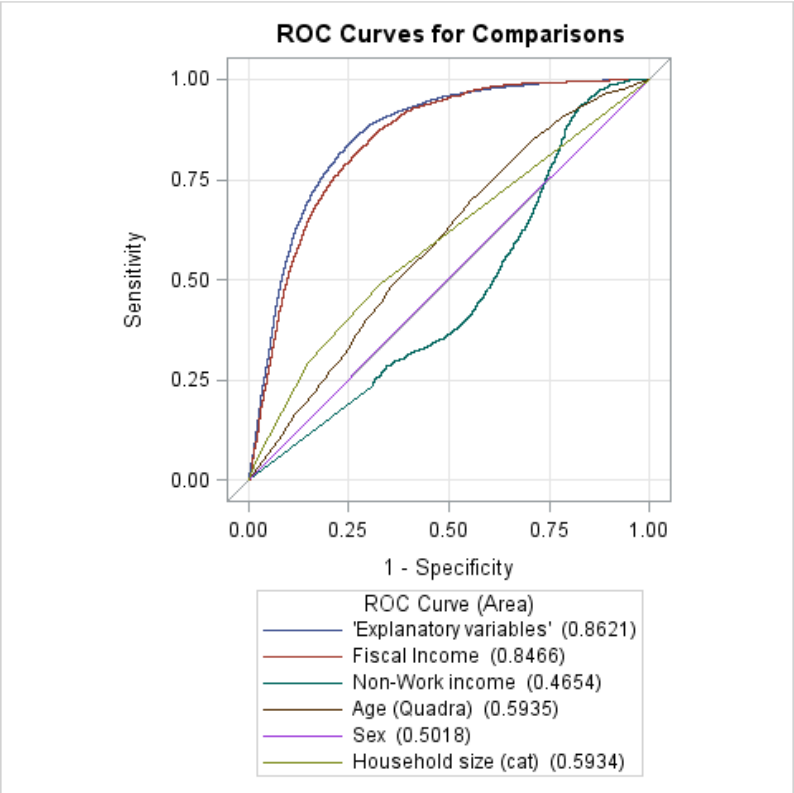


Figure 25: Explanatory power of auxiliary variables for AROPE rate (SILC 2016)

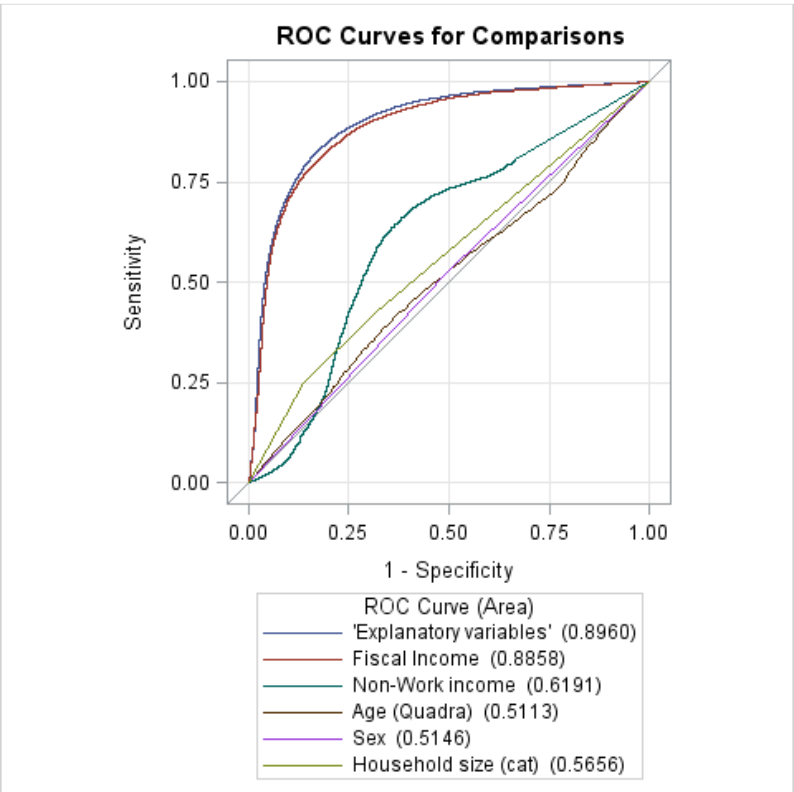


Figure 26: Effect of NUTS-2 variables on the model for AROP rate (SILC 2016)

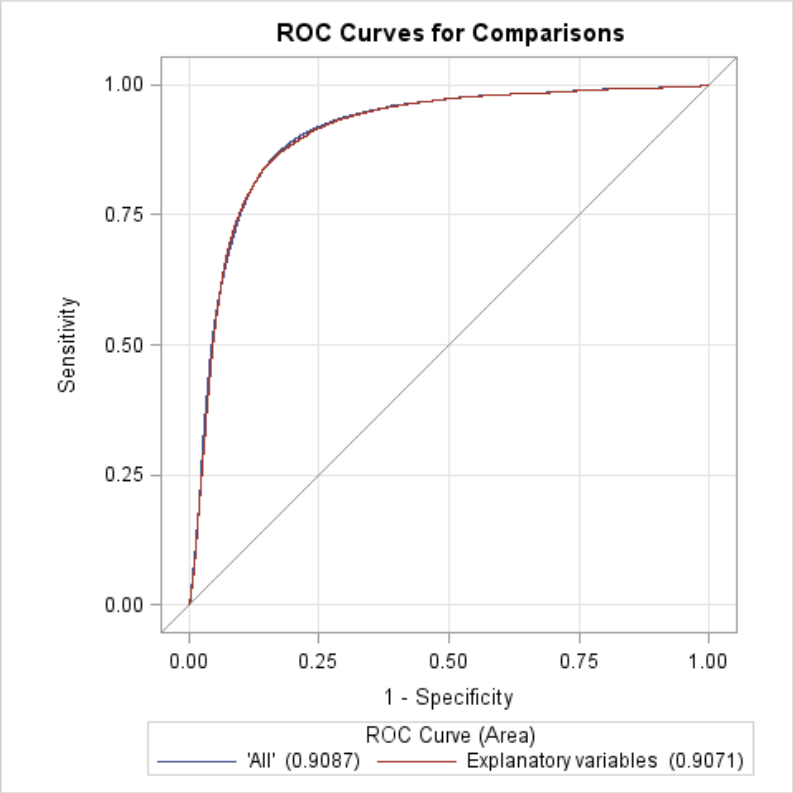


Figure 27: Effect of NUTS-2 variables on the model for LWI rate (SILC 2016)

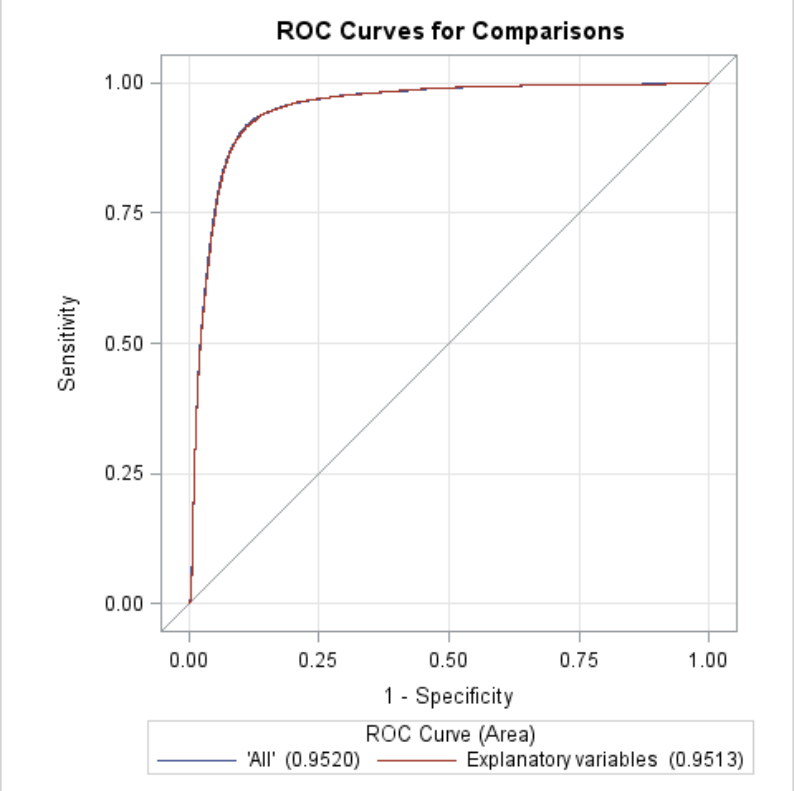


Figure 28: Effect of NUTS-2 variables on the model for SMD rate (SILC 2016)

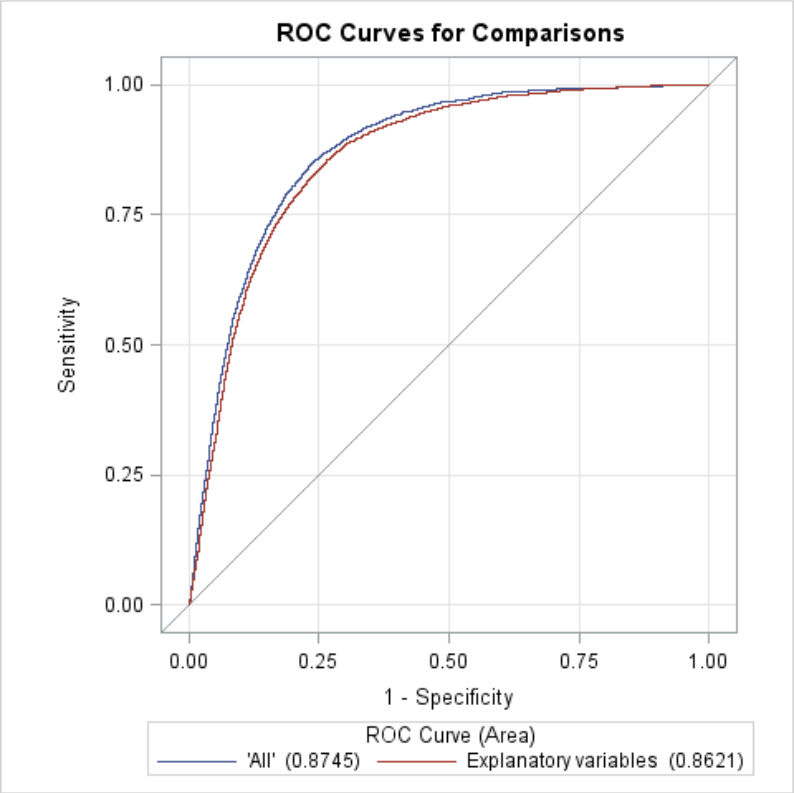


Figure 29: Effect of NUTS-2 variables on the model for AROPE rate (SILC 2016)